AD_____

Award Number DAMD17-94-J-4270

TITLE: A New Generic Method for the Production of Protein-Based Inhibitors of Proteins Involved in Cancer Metastasis

PRINCIPAL INVESTIGATOR: Marshall H. Edgell, Ph.D.

CONTRACTING ORGANIZATION: University of North Carolina, Chapel Hill
Chapel Hill, North Carolina 27599-4100

REPORT DATE: August 1999

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	August 1999	Final (1 Aug 94 - 31 Jul 99)

4. TITLE AND SUBTITLE
A New Generic Method for the Production of Protein-Based Inhibitors of Proteins Involved in Cancer Metastasis

5. FUNDING NUMBERS
DAMD17-94-J-4270

6. AUTHOR(S)
Marshall H. Edgell, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of North Carolina, Chapel Hill
Chapel Hill, North Carolina 27599-4100

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Our objective was to develop a method to make protein-based inhibitors against protein targets and as the test case to a proteinase involved in metastasis, stromelysin. The approach used phage display and a display framework which was expected to bind preferentially to the active site pocket of target proteins. While we were able to find peptides in phage display libraries that bound to active stromelysin or to cadmium inactivated stromelysin and could find binders to a test target, papain, in a constrained loop library we were unable to get the fully constrained scaffold protein to bind to its cognate target, subtilisin in the phage display system. Hence the overall objective was not obtained. As part of our efforts to characterize the binding epitope display framework, eglin c, we did develop a new method for using mutagenesis to study protein structure which we call patterned library analysis. We carried out a proof-of-principle experiment using the new method in which we were able to reproduce the values for α-helix propensity indicating that the method does indeed work. This method should have broad utility as a method for the assessment of hypotheses concerning the determinants of protein structure.

14. SUBJECT TERMS
Breast Cancer
stromelysin, phage display, eglin c, combinatorial libraries

15. NUMBER OF PAGES
94

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

MHE ____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

MHE ____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

MH Edgell 9/9/99

PI - Signature Date

Table of Contents

INTRODUCTION

Our long range interests were to design an array of protein-based inhibitors for proteins involved in various steps in metastasis. The primary objective of this project was to test the viability of a new approach for designing inhibitors for proteins and as 'proof-of-concept' to generate an inhibitor for stromelysin, a proteinase implicated in metastasis. Our approach was to use protein engineering to redirect the activity of a naturally occurring proteinase inhibitor, eglin c to the targets of interest. Since the state-of-the-art of protein structure/function is such that it would be unlikely that we could design an effective inhibitor in a single pass, we planned to use genetic screening techniques to find the highest affinity variants in large libraries of structural variants centered around the basic design concept. Our approach used phage display for genetic screening and a display framework which was expected to bind preferentially to the active site pocket of target proteins and to provide the constraints necessary for high-affinity binding. Hence the project consisted of:

1. getting the phage display system working in our laboratory.
2. producing stromelysin for use as a binding target
3. getting our protein scaffold to work in phage display
4. determining how to get binding to proteinases that can cleave the binding molecules
5. building suitable libraries for binding to stromelysin
6. finding binders to stromelysin

We were able to accomplish steps 1,2,4,5 and 6 but not 3. We were able to get all of the parts of the project to work except the crucial one involving the scaffold protein that was intended to provide the constraints necessary for high-affinity binding. Early in the project, before we had become familiar with some of the eccentricities of the phage display system, we obtained evidence that we interpreted as positive concerning the utility of our scaffold protein. Late in the project we discovered that the scaffold protein did not, in fact, provide the display function necessary for phage display. This made it impossible for us to generate high-affinity inhibitors using the approach envisioned in this project.

We are disappointed that this project has not made a direct contribution to the attack on breast cancer as was the project goal. However, as part of our efforts to characterize the binding epitope display framework protein, eglin c, we did develop a new method for using mutagenesis to study protein structure which we call patterned library analysis. We carried out a proof-of-principle experiment using the new method in which we were able to reproduce known values for α-helix propensity indicating that the method does indeed work. We expect the method to be used in various approaches to exploring the determinants of protein structure. This should make a contribution to protein structure prediction, one of the outstanding unsolved problems in biology. Like most basic science tools, that should, in the long term, enhance our capacity to deal more effectively with diseases such as cancer.

BODY

Summary

Initial Project Goals

1. Convert eglin c into a framework for the development of protein-based protein inhibitors.
2. Incorporate a binding epitope for stromelysin into the inhibitor framework and identify stromelysin binders in libraries designed to optimize binding to the desired target.
3. Test the stromelysin binding eglin variant for its capacity to inhibit metastasis.

Failures

1. The main project failed ultimately in step 1 although we did get some stromelysin binders by using simplified scaffolds (step 2).
    a. During the first year we discovered that wild type eglin did not work in the phage display system which was central to our project. We assumed that this was due to the site on eglin to which it was fused to the phage attachment protein, pIII.
    b. Consequently we constructed a circularly permuted version of eglin variant (peglin) to move the N-terminus to the opposite side of eglin to overcome this problem.
    c. Control experiments indicated that this modified form of eglin did work in the phage display system.
    d. A large number of experiments were then done to improve our use of the phage display system, to enhance library construction, to make stromelysin, and to document that binders could be found to target proteinases. We found binders to various enzymes using randomized peptides and to stromelysin using a simplified scaffold consisting of an18 amino acid loop.
    e. When we began to use peglin under phage display conditions that were now known to be reliable we were not able to get peglin to bind to its cognate enzyme.
    f. After considerable frustration we carried out a reassessment at the beginning of the last year of the project as to whether peglin would function at all as a phage display scaffold and discovered that it did not! As this was very near the end of the grant period we did not worked out the reason for this problem although we understand the source of the difference in our current results and the earlier ones. Instead we used the remaining time to explore a successful ancillary project that had arisen during the main project.

Accomplishments

1. We carried out 'proof-of-principle' experiments for a new method to utilize mutagenesis and combinatorial libraries to assess hypotheses concerning the determinants of protein structure. This project grew out of our efforts to characterize the model protein that we have been using as a scaffold.
    a. As proof of principle we used our new method to show that it could reproduce known values for the helix propensities of amino acids (ref. 3).
    b. We feel that the development of this new method is a very significant one in that we expect it will be used broadly to enhance protein structure prediction algorithms. Protein structure prediction is, of course, one of the major outstanding problems in biology.

2. Characterization of eglin c.
    a. We have determined several of the basic thermodynamic properties of eglin c and shown that these apply to his-tagged eglin c as well as to wild type (ref. 1).
    b. We have uncovered a non-ideality between the native and denatured states of eglin c that explains the differences between van't Hoff and calorimetric denaturation enthalpies. This observation may, in fact, apply to other proteins whose behavior is otherwise consistent with a two-state mechanism for unfolding (ref. 2).

Objectives of the Project

Our long range interests were to design an array of protein-based inhibitors for proteins involved in various steps in metastasis. The primary objective of this project was to test the viability of a new approach for designing inhibitors for proteins and as 'proof of concept' to generate an inhibitor for stromelysin, a proteinase implicated in metastasis. Our approach was to use protein engineering to redirect the activity of a naturally occurring proteinase inhibitor, eglin c to the targets of interest. To redirect the activity of eglin c the plan was to retain the native constraints in the framework that reduce the number of non-productive conformations accessible to the binding epitope but to replace the native binding epitope with one for stromelysin. Since the state of the art of protein structure/function is such that it would be unlikely that we could design an effective inhibitor in a single pass, we planned to use genetic screening techniques to find the highest affinity variants in large libraries of structural variants centered around the basic design concept. The plan was to employ multiple cycles of design, construction, affinity screening, and biophysical analyses. Once high affinity ($K_d > 10^{-9}$ **M**) stromelysin inhibitors were identified we planned to test their effects in standard assays for metastasis.

Investigations.

Most of the investigations described below were done in parallel. They are reported here as isolated projects to provide a more coherent report. Hence it is important to keep in mind when reading a section that we did not necessarily know at the time we did those experiments everything that has been described in earlier sections of the report. The main conclusion of section I in which we show that the basic premise of our project was flawed, at least in terms of the model protein chosen, was not found out until the last year of the project. We describe how our early results (mid 1996) although correct and reproducible misled us in terms of the properties of our model protein. On August 23, 1998 we collected data that indicated that the scaffold protein chosen was not going to work in the context we were trying. From that point on we focused on a scientifically promising lead that had developed in our laboratory while doing basic characterization of the eglin c scaffold protein. This ancillary project has been successful and we hope that it will provide a new tool to help in the characterization of proteins and the determinants of protein structure.

We are disappointed that this project has not made a direct contribution to the attack on breast cancer as was the project goal. However, the successful components of the project have led to a new method with broad basic science implications. We expect the method to be used in various approaches to exploring the determinants of protein structure. This should ultimately enhance our capacity for protein structure prediction. Like most basic science tools, that should, in the long term, accelerate learning about cancer which will ultimately increase our capacity to deal more effectively with the disease.

<u>I. Eglin c as a Phage Display Scaffold.</u>

Our basic approach was to extend the reach of traditional protein engineering by constructing large libraries of structural variants around a central design concept and then using the phage display system to screen the population for binders. Hence it came as a considerable shock when we discovered that wild-type eglin c, when fused to the M13 gene III protein in the M13 particle, does not bind to a target to which the free inhibitor

normally binds (e.g. subtilisin). We presumed that this problem was due to the fact that the eglin c molecule attaches to the phage via it's C-terminus which is close to the active site and that this blocks access to the eglin c binding epitope (see Figure 1). That is, the eglin c C-terminus appears to be too close to the loop containing the residues that bind with the target proteinase.

To move the gene III fusion point further away from the binding epitope we made eglin c variants with various C-terminal truncations and made the M13 gene III fusions via linkers of 5 and 9 prolines. None of these eglin c variants bound to subtilisin as measured by enrichment in a phage display assay.

Peglin Construction. We then made a circularly permuted version of eglin c in which the C-terminus was moved to the side of the protein opposite from the residues that bind with the target. We designed an eglin variant in which the wild-type N and C termini have been joined together and new termini created by opening up a tight turn on the opposite side of the protein (Figure 1). This involved various modeling exercises to evaluate residues for a tight turn and at exactly which point to truncate the existing N and C termini. A construction was carried out and was verified by sequencing. We then tested the new construct for its ability to inhibit subtilisin and function in phage display.

Peglin is Active Against Subtilisin. The peglin construct, in the plasmid pET28, was transformed into E. coli BLR*lysS* and grown up in 2YT. Cell lysates were prepared by resuspending a frozen cell pellet in 50 mM Tris pH 8.0 and treatment with 100 ug/ml of lysozyme. Samples boiled for 10 minutes in 1% SDS were then examined on 15% acrylamide gels (Figure 2). This indicated that the protein was relatively stable since if it were not the case there would be less peglin than eglin in the lysates. Peglin had a specific activity in these lysates essentially the same as eglin; relative specific activity of 1.05 as measured by its capacity to inhibit subtilisin. This tells us that the circularly permuted version of eglin does indeed bind to its normal target, subtilisin. It does not tell us whether peglin will function in the phage display system.

Initial Evidence that Peglin Binds in the Phage Display System (~May 1996). The gene for the circularly permuted version of eglin was then transferred to the M13 phage used for phage display. That is, the peglin gene was fused to the N-terminus of the M13 pIII gene. To test for binding we used a target proteinase binding assay. Samples are applied to wells in a 96-well plate coated with subtilisin and the amount of phage that binds to the wells is assessed with alkaline phosphatase conjugated anti-M13 antibody. We grew up liter cultures of the M13 phage containing the peglin fusion product and mixed phage with subtilisin coated wells to see if it would bind. To get a decent molar ratio of phage to enzyme we concentrated the phage using polyethylene glycol precipitation. Peglin fusion phage were then applied to subtilisin coated wells and the amount of binding assessed using alkaline phosphatase labeled antibody against M13. Various dilutions of the concentrated phage stock were employed (Figure 3). As a control for this experiment equal numbers of phage expressing a phage display epitope for strepavidin was applied to wells coated with subtilisin (Figure 3). 1.4 times as many peglin phage bound to subtilisin coated wells as did control phage. This weak binding was attributed to the presence of phosphate in the binding media which reduces subtilisin activity.

To further assess peglin binding to subtilisin we carried out an enrichment assay. In such an assay we mixed a small quantity of peglin with a large population of non-binding phage and subjected the population to several rounds of panning against subtilisin. One then looks for a change in the input ratio. This can be done

easily since the peglin phage contains a gene for the alpha fragment of beta-galactosidase and hence will make blue plaques on indicator plates. The non-binding phage do not carry the alpha fragment and hence make white plaques. Wells in a 96-well plate were coated by exposure to solutions of subtilisin at 25 ug/ml and incubation overnight. The wells were washed 5 times with 50 mM Tris pH 8.0 and then treated for 12 hours with 1% BSA to block any remaining non-specific binding sites in the wells. Phage were added to the wells and allowed to bind for 4 hours. The wells were then washed eight times with PBS-0.1% Tween 20 and then the bound phage were eluted with 50 mM glycine-HCl pH 2.0. After neutralization using 200 mM $NaPO_4$ pH 7.5 the eluted phage were titered. This assay gave an enrichment of anywhere from 7 to 466 fold (Table 1).

These experiments convinced us that peglin could bind to subtilisin, the normal target for the eglin epitope and hence we proceeded with a series of experiments over the next two and a half years to transfer other epitopes to the peglin scaffold. The problems with this interpretation are delineated in the section below on Final Evidence that Peglin Really Does Not Bind to Subtilisin (August 1998).

Enhancements to the Peglin Phage Display Vector. I. Restriction Sites. Construction of libraries seemed to take much more time than was convenient for this project that was so dependent on library construction. It seemed to us that the problem was that each library was an idiosyncratic affair involving different restriction sites. Each new restriction site employed seemed to take a large amount of time for us to optimize the conditions for effective cleavage while leaving ligatable ends. We decided that it would be worth the effort to build a vector that would use the same restriction enzymes independent of where we intended to make the mutations in the parent gene. To do this we intended to use a restriction enzyme such as Ear I that cleaves outside of the enzyme recognition sequence. If one uses PCR primers that have such a site at their ends the PCR product can be cleaved to leave a ligatable ends inside of the template sequence at any position of choice independent of any restriction sites in the template (Figure 4). This also means that one can use the same restriction enzyme for any library. Optimization for one library should carry over to all of those that follow.

To make this work we needed a vector that had no Ear I sites. Our original peglin containing M13 phage derivative had two Ear I sites. These were removed by site directed mutagenesis.

Enhancements to the Peglin Phage Display Vector. II. White Plaque Variant. One of the methods to find rare phage which bind to a target of interest is to look for enrichment of the variants containing potential binding epitopes relative to non-binder phage. One way to measure enrichment is to be able to distinguish library phage from non-binder phage based on the color of plaques they make. Our library vectors all made blue plaques on indicator plates. While we had several non-binder phage that gave white plaques on indicator plates they all produced many more phage progeny per generation than did the mBAX phage into which peglin was inserted. We decided to prepare a white plaque variant of the mBAX vector, that is, the base vector without peglin. This was done using PCR primers to amplify up all of the mBAX phage except for a 200 bp deletion within the beta-galactosidase alpha fragment (Figure 5).

Libraries Involving the Peglin Scaffold. The first peglin based library was constructed by removing an Xho/Xba fragment from peglin and replacing it with a degenerate PCR product generated from degenerate oligonucleotide primers. This library had eight residues in the center of the binding epitope loop (Figure 1) randomized and replaced two arginines with alanine. The two arginines in the native molecule interact with

residues in the loop (an aspartic acid and a threonine). Since those loop residues have been randomized we chose to remove the arginines. The diversity of this library was low, about 3 x 10^5 different phage variants. This library was never tested against stromelysin since at the time we made the library we did not have any stromelysin and by the time we had obtained a useful supply of stromelysin we had learned that the peglin scaffold was not working in the phage display system.

Final Evidence that Peglin Really Does Not Bind to Subtilisin (August 1998). Having decided initially that peglin fit our preconceptions and was indeed a suitable scaffold protein for displaying binding epitopes for novel targets we set about learning how to use the phage display system effectively, to working out methods to enhance the library construction process, to making suitable quantities of stromelysin, and to document that binders could be found to target proteinases. Once the phage display system was under control and we began to use peglin under conditions that were now known to be reliable we discovered that could not get peglin to bind to its cognate enzyme. Hence we were forced to reassess the basic premise that peglin was a suitable scaffold for phage display. That meant reexamining whether the unmodified peglin protein with the subtilisin epitope would bind to subtilisin.

We retested whether peglin would inhibit subtilisin. One possibility for our lack of success at getting binders was that the peglin strain had mutated and become inactive, so we made several new phage preparations from old isolates. Large cultures were grown up to produce enough phage to be able to detect inhibition. When these polyethylene glycol concentrated stocks were mixed in various dilutions with subtilisin the peglin stock inhibited subtilisin more completely than did the non-binder phage stock just as in the earlier experiments (Figure 3). However, now looking more intensely for a problem with these results, I realized that the control non-binder phage produced a much higher concentration of phage in the growth cultures than the peglin phage and hence when comparing equal quantities of phage (PFU/ml) we were adding much different volumes of the two concentrated phage stocks. The experiment was then repeated using equal volumes of phage stocks instead of equal amounts of phage. In that experiment we discovered that equal volumes of the phage stocks gave exactly the same amount of inhibition (Figure 6) even though the amounts of phage in equal volumes differed by a factor of 100! In particular, look at Figure 6, rows C,D, and E. These different phage preparations behaved exactly the same in terms of inhibition even though the amount of putative inhibitor, that is, phage, were quite different in the preparations. The implication of that was that the inhibition was due to a contaminant in the concentrated phage stocks, most likely polyethelene glycol, rather than the phage itself. This was repeated by two different people in the lab and with several different peglin isolates.

Another test done to assess peglin binding to subtilisin was to take phage preparations and see whether they bound to subtilisin adhering to wells in 96-well plates. In this case binding was assessed by using alkaline phosphatase conjugated antibody against M13 to measure the amount of phage bound to the well. In this test we compared the amount of phage binding to wells containing subtilisin with wells containing catalase to control for non-specific binding. Testing ten different peglin isolates we found no binding to subtilisin that was higher than the non-specific control (Figure 7). Evidence for binding by one of the peglin isolates would be larger numbers in any given column in rows A-D (the subtilisin coated wells) versus E-H (the catalase coated wells). None of the isolates gave such a result.

These experiments indicated that peglin did not bind to subtilisin. The other piece of data originally indicating that peglin bound to subtilisin was an enrichment experiment in which blue plaque producing peglin phage were mixed with white plaque producing non-binder control phage and the mixture panned against a subtilisin coated well. As we became more experienced with phage display in the two years since those initial experiments we learned that single round panning experiments were unreliable. They often gave irreproducible results. In addition, the single round competition experiments were very sensitive to the relative burst sizes of the two phage and this added to the problems with a single round panning experiment as an assay for binding. As a consequence we often obtained results from single round panning experiments that did not stand up on reproduction. That was a gradual realization and it did not cause us to question our original conclusions about peglin since we thought we had other convincing data that peglin would bind to subtilisin.

These new experiments concerning peglin's ability to bind to subtilisin were done in August of 1998. We concluded at that time from the data from those experiments that peglin was not going to work as a scaffold for creating new protein-based inhibitors. Our laboratory efforts for the remaining 12 months were then turned to a scientifically promising lead that had developed from the ancillary project involving the biophysical characterization of the base scaffold protein, eglin c. That project has been quite successful and is described in section V below.

II. Truncated Eglin as a Scaffold.

It has been shown by others that a truncated form of eglin inhibits it's normal targets with the same efficiency as does the wild type eglin protein. So, while circular peptides would not make good physiological inhibitors due to their susceptibility to proteolytic degradation they could serve as a scaffold for the discovery of binding epitopes that could then be moved to the full circularly permuted form of eglin, peglin. So, in parallel with our efforts to develop the circularly permuted form of eglin as a scaffold we began to explore the utility of this truncated form of eglin as a scaffold to search for binding epitopes targeted to other proteins. We called the truncated form teglin. Teglin consists of an eighteen amino acid circular peptide closed via a cysteine bond and fused to the pIII M13 protein for display via a linker of QGGGG (Figure 8). We tested teglin for binding to subtilisin by both the inhibition assay and single cycle enrichment. Polyethelene glycol concentrated teglin phage inhibited subtilisin. We got no enrichment from a single cycle. At the time we interpreted this as evidence that teglin worked and that enrichment failed due to the anticipated very tight binding of teglin to subtilisin. Knowing what we know now I can't say whether or not teglin inhibited subtilisin. As the next section will indicate we do know that teglin does work in phage display as would be expected from experiments that others have done with constrained loops and phage display.

Conversion of Teglin to a Papain Binder. A. The Papain Cleavage Epitope. This project was a methods development project. Our ultimate target for inhibitor development was stromelysin, a proteinase implicated in metastasis. However, to use stromelysin as a target we needed to make significant quantities of the protein. That was being pursued in a parallel activity and is described below (section IV). To work out methods we used a commercially available enzyme, papain, as a target proteinase. We chose papain since it is cheap, it's structure is known, it is a proteinase from a different class of enzymes (cysteine proteinase) than

subtilisin (serine proteinase) and it's cleavage specificity is known. We also wanted to work with a proteinase since there are serious issues to work out for using phage display to find protein binders to a proteinase. One could expect background binding of proteins to a proteinase. One could expect that intermediate level binders might be cleaved by the proteinase and released from the affinity matrix preventing enrichment.

The binding epitope in teglin was then replaced by a sequence that is cleaved by papain (Figure 9). This construction was verified by sequencing. Two isolates of the modified teglin display phage (white plaque producer) were then mixed with an M13 phage, m663, displaying a peptide for a non-cognate target that produces blue plaques on indicator plates. These mixtures were tested for enrichment in wells of a 96-well plate coated with either active or papain inactivated by treatment with Mg^{+2}. To control for unspecific binding and enrichment we included a teglin/m663 mixture and a mixture of m663 with an M13 without any display peptide. We found about a ten fold greater enrichment for the teglin variants with the papain binding epitope relative to the m663 phage displaying a peptide for a non-cognate target (Table 2). On the other hand teglin with the subtilisin binding epitope was enriched 3-5 fold.

Conversion of Teglin to a Papain Binder. B. Papain Cleavage Epitope Library. A library of binding epitopes in the teglin scaffold was constructed (Figure10) by insertion of an Xho/Xba fragment to replace the subtilsin binding epitope. The new fragment maintained the papain cleavage sequence, phe-ala, and randomized four residues, one just upstream of the cleavage site and three downstream. The library was subjected to four rounds of enrichment in wells coated with inactivated papain. Thirty-four phage isolates from the enriched population were then tested by a single round of competition against m663 a blue plaque producer on X-gal indicator plates. Two of the phage isolates had considerable better enrichment than the rest (Table 3).

We carried out another set of panning with the library against inactivated papain. This time we used six rounds of panning with each round being followed by amplification to increase the diversity of the phage at each round. We had discovered that only a few thousand phage were getting through the first two rounds of panning prior to amplification. This was nice for getting only a few phage types out of the panning process but not good for trying to find all of the different types of binders. After the panning 90 isolates were picked and tested for binding to inactivated papain using alkaline phosphatase conjugated anti-M13 antibody to assess binding (Figure 11). These panned phage divided up into three groups, 21 non-binders, 11 intermediate binders, and 58 high binders. Eight isolates were sequenced (Figure 12). All five from the high binder group had undergone some mutation that eliminated the cysteine link to form the constrained loop. Only two of the eight retained the cysteine constraint.

This result with papain as the binding target had both positive and negative interpretations. On the positive side we had been able to retarget the eglin binding loop to a non-cognate proteinase target. We were able to see binding to a proteinase that might have cleaved the binding epitope preventing enrichment. On the down side most of the binders had lost the cysteine which was designed to hold the binding loop closed. This converted the binding sequence into a free peptide. The trouble with free peptide binders is that there is an upper limit on the binding potential for peptides imposed by the large entropy penalty paid from the many degrees of freedom unbound state to the reduced degrees of freedom bound state. This upper limit for the

binding of free peptides is below what is usually thought to be necessary for getting enough binding under physiological conditions to produce an effect.

Construction of General Libraries Using the Teglin Scaffold. Despite the downsides, the papain results seemed promising to us and so we constructed two general libraries that we thought could be used to look for stromelysin binders. Initially we had imagined using libraries like the one for papain but which incorporated the stromelysin cleavage sequence. However, results using randomized peptides that will be discussed below suggested to us that less specific libraries might be a more effective use of our time and of course less specific libraries would certainly have more utility for use with different targets.

In wild-type eglin there are two interactions between residues in the binding loop and the underlying beta-strands that increase the affinity of binding. We made one library that replaced the arginines in the beta-strand with alanines and another library that put randomized codons at those sites. In the latter library the idea was to provide an opportunity for different binding loop/beta-strand interactions to form. In both libraries we randomized eight residues centered on what would be the scissile bond in a cleavable peptide. The first library, 8X+2A, had a complexity of 2.3×10^8. The second library, 8X+2X, had a complexity of 1.5×10^9. These libraries had quite good diversity which we attribute to benefits of using the Ear I strategy and optimized vector described in a section above (Enhancements to the Peglin Phage Display Vector. I. Restriction Sites) which were also applied to the teglin vector.

Utilization of Teglin Libraries. We began to use these libraries against various targets, a bank of model targets (cheap, commercially available enzymes), stromelysin and subtilisin. As the binding controls for these experiments we added the circularly permuted version of eglin that is supposed to bind to subtilisin to a catalase binder we had isolated and a strepavidin binder. In contrast to the other binding controls, the peglin control did bind to its cognate target, subtilisin. This led to more and more stringent testing of the peglin phage as described above and the ultimate decision that the peglin scaffold was not going to work in the phage display system. As a consequence further exploration of the truncated versions of eglin were abandoned since they were intended to serve primarily as a way to explore the binding system and identify potential binding epitopes for use with peglin.

III. Exploring Search Strategies for Binders Using Randomized Peptides. The phage display system was initially developed by George Smith using randomized peptides as the source of potential binding epitopes. We began our exploration of the phage display binding system using an existing library of randomized peptides since we could begin without any recombinant DNA constructions being necessary and because that system has the most information about it already in the literature.

Increasing Diversity Prior to Amplification. The initial protocol for doing phage display, taken from a phage display workshop manual, involved two rounds of panning followed by an amplification step. Adding 100 ul of a phage stock, usually at 10^{11} PFU/ml, to a microtiter plate well for panning in the first step often resulted in only a few hundred phage at best entering the amplification step. As a consequence we modified our protocol to incorporate an amplification step after each panning step and to utilize 12 microtiter plate wells per

sample instead of only one. This increased the number of phage entering that first most stringent amplification step to ~10^4 PFU.

We also tried using 'Immunosorb' tubes (Nunc, Naperville, IL) in that they had a much larger surface area and the expectation was that more phage would be carried through the panning cycles. In contrast to this expectation in control experiments very few phage bound.

The Assorted Enzyme Set. To provide a series of targets to use to work out methodology we selected eight enzymes on the basis of commercial availability, cost and potential utility of an inhibitor. The enzymes selected were: alcohol dehydrogenase, aldolase, apha amylase, catalase, enolase, hexokinase, L-lactate dehydrogenase, and ribonuclease. Using the library of randomized peptides and six rounds of panning and amplification we found binders to four of the eight enzymes that bound with a signal of at least four times background (Table 4). The binders to ribonuclease are potentially of some interest since ribonuclease inhibitors are valuable laboratory reagents. However, no further work was done with those binders.

Random Peptide Binders to Stromelysin. We screened the randomized peptide library against stromelysin using the high diversity protocol of six rounds of panning and amplification. Our expectation was that weak binders might be cleaved by the enzyme and hence would not bind. If tight binders were not cleaved and existed in the initial population that would be OK since we are primarily interested in tight binders. However, we anticipated that there might not be any tight binders and that we would need to rescue weak binders and then find ways to increase the binding affinity by further mutagenesis. To try to capture weak binders we tested as targets both native stromelysin and stromelysin inactivated by treatment with $MgCl_2$ or $CdCl_2$ to displace the Zn^{+2} required for activity in native stromelysin. Binders were found to all three forms of stromelysin (Figures 13,14,15).

We sequenced thirty clones that bound to the cadmium treated stromelysin. Four different sequences were found in the thirty isolates (Figure 16). Alignments made around the amino acids common to all of the may define a stromelysin binding motif (Figure 17). We has anticipated that we would move these sequences into the full scaffold when that became available but since the full scaffold (peglin) was discovered, very late in the project, to not be active in the phage display system that was never done.

IV. Production of the Target Protein Stromelysin

Summary of Stromelysin Production. The objective of this project was to develop a methodology for creating protein-based inhibitors against proteins of interest. The prime target was stromelysin since that protein has been implicated in the metastasis of cancer. Being a proteinase stromelysin we expected that it would be able to accommodate a redesigned proteinase inhibitor in its active site. When we began the project there was no commercial source for this enzyme so we expected to have to build an expression system to produce it. We were able to acquire a small sample of active stromelysin and of prostomelysin at the beginning of the project from Roche Biosciences to validate our assays. After considerable effort making recombinant constructions and working out conditions to produce active stromelysin from prostromelysin we obtained a gift of a few milligrams of mature stromelysin from Parke-Davis. This was a sufficient amount for a large number

of phage display experiments so at that time we discontinued our work on producing our own source of stromelysin.

Activity Assays for Stromelysin. We obtained a gift of the mature form of stromelysin from Dr. Paul Cannon at Roche Biosciences in Palo Alto, CA. This material was diluted to a working concentration of 1 mg/ml in 25 mM Tris, pH 7.25, 10 mM $CaCl_2$, 0.05% Brij-35 and stored at -70C in small aliquots. For our stromelysin assay we used an activity assay designed for vertebrate collagenase. That assay utilized the hydrolysis of a thiopeptide substrate, Ac-pro-leu-gly-[2-mercapto-4-methylk-pentanoyl]-leu-gly-OEt purchased from Bachem Biosicences (Philadelphia, PA) and Ellman's reagent (DNTB or 5.5'-dithio-bis(2-nitrobenzoic acid) purchased from Sigma Biochemical CO. (St. Louis, MO). The assay conditions are 50 mM MES pH 6.0, 10 mM $CaCL_2$, 106 mM thiopeptide substrate and 1 mM DTNB in a 100 ul final volume. Thiopeptide was prepared at 76 mM in 80% acetic acid. DTNB was prepared at 20 mM in 95% EtOH. MES was prepared as a 1 M solution, filter sterilized and stored at -20C in 10 ml aliquots. CalCL2 was prepared as a 1 M solution and autoclaved. A microtiter plate reader (Molecular Devices) was used to monitor activity as indicated by change in adsorbance at 405 nm at room temperature for 30 minutes with 11 second intervals between readings. This assay has a sensitivity of about 0.5 ug/ml and a maximum of about 10-12 ug/ml (Figure 18).

Activating Prostromelysin. The stromelysin enzyme is synthesized as a pro-enzyme. When we began this project it had been recently discovered that the pro sequences of several proteinases were involved in the proper folding of the enzymes. For those enzymes synthesis of the catalytic core without the pro sequence gave mostly inactive protein. It also seemed likely that expression of an active form of a proteinase in *E. coli* would be toxic for the bacterium. Hence we assumed that we would have to make stromelysin from prostromelysin and hence would to verify that we could convert prostomelysin to mature stromelysin. As it turns out a gene for the mature form of stromelysin does produce active enzyme and it is not prohibitively toxic in *E. coli* but that was not known at the time and so the following work was done.

Dr. Paul Cannon also provided us with a sample of full length human prostromelysin-1 purified from IL-1 stimulated human gingival fibroblast conditioned medium. The material is stored at a concentration of 0.8 mg/ml in 50 mM Tris pH 7.4, 0.2 mM NaCl, 5 mM CaCl2, and 0.02% NaN3 at -70C in 100 ul aliquots. A trypsin method was used for activation using the assay described above to measure resultant stromelysin activity. Trypsin processes the 58kD prostromelysin to a 45 kD active form which partially autoprocesses to smaller active forms around 28 kD. Trypsin does not act on the substrate to produce color. Approximately 2-3% of the prostromelysin was activated in this experiment (Figure 19). Longer times did not increase the yield. A wider range of trypsin concentrations was tried using a gel assay to monitor conversion to a shorter form since this assays uses much less material (Figure 20). Prostromelysin was treated with 100, 25, and 4.16 uM trypsin and assayed for the production of 48 and 28 kD forms of stromelysin. Treatment with 25 uM trypsin led to conversion within 10 minutes. Treatment for 45 minutes with trypsin at any of the test concentrations led to complete degradation of the prostromelysin. The best activation treatment within the range tested was 4.16 uM trypsin for 10 minutes. Using the activity assay we found that this treatment converted approximately 9% of the prostromelysin to the active form (Figure 21).

Clones for Prostromelysin. Several clones for prostromelysin were constructed using PCR to amplify various portions of the stromelysin-1 gene which was obtained from Dr. Lynn Matrycian. Different clones were constructed since we did not know the exact boundaries of sequence that could be expressed in E. coli and would produce a properly folded protein. Amplified DNA sequence was cloned into both pET3d and pET28a expression vectors (Novagen, Inc. Madison, WS). pET3d requires a T7 promoter for expression and hence provides very tight expression control. pET28a is driven by a T7/lac promoter and adds a his-tag to the cloned prostromelysin gene. Neither promoter is active in bacteria that do not have a source of T7 RNA polymerase. Constructions are maintained in hosts without the T7 polymerase gene since high levels of expression are usually growth inhibitory and hence lead to growth advantages for variants that have lost either the protein sequences or the expression machinery. Our plasmid constructs were transformed into BLR (DES) pLysS, a bacterial host expressing t7 RNA polymerase under tight control. Five independent transformation isolates were tested for activatable prostromelysin. Cultures were grown in LB at 37C with shaking at 250 rpm to an OD_{600} between 0.6 and 0.9. Expression was induced by adding IPTG to 0.4 mM for the 3d hosts and 1.0 mM for the 28a hosts and the cultures were grown for an additional 3 hrs. The cultures were chilled on ice for 5 minutes and the cells collected by pelleting at 4C. The pellets were washed in 0.25 growth volumes of 50 mM Tris pH 8.0, 10 mM $CaCl_2$. The cells were pelleted again and frozen at -70C. To lyse the cells the pellets were thawed in a water/ice bath, resuspended in cold 1/10 growth volume 50 mM Tris, pH8.0, 10 mM $CaCl_2$ and lysozyme was added to 0.2 mg/ml. After 20 minutes on ice the material was sonicated at 90%intermediate output (Branson sonifier) for 1 minute. This lysate was cleared by centrifugation in an Eppendorf micocentrifuge. The supernatant was collected and stored at 4C. Total protein was determined using the Pierce BCA method.

Two of the five clones tested, strom3 and strom9, produced stromelysin activity after activation with trypsin (Figure 22).

Purification of his-tagged prostromelysin. We chose to work up the strom9 clone since it has a his-tag to facilitate purification while the strom3 clone does not. Duplicate cultures of bacteria containing the strom9 expresser plasmid were grown up and induced as described above. Samples were collected every hour for 6 hours after induction. Lysates were prepared as described above and analyzed by electrophoresis on polyacrylamide gels (Figure 23). On induction a band at 33 kD is seen to increase in intensity with time of induction. The predicted size for this prostromelysin product is 32 kD as the clones were intentionally truncated during construction to make the shortest possible 'pro' form of stromelysin.

To purify the soluble his-tagged prostromelysin from these cells a frozen pellet from 1250 ml of culture was thawed on ice and resuspended in 10 ml of dHOH and the final solution made up to 0.2mg/ml in lysozyme, 10 ug/ml in DNase, 10 mM $CaCl_2$, and 1 mM PMSF. After 30 minutes on ice the solution was sonicated (5-30 seconds bursts with 30 seconds of cooling between each). The solution was then cleared by centrifugation at 4C. The supernatant was filtered through a 0.45 micron syringe filter and kept on ice while a 4 ml his-bind resin (Novagen, Inc. Madison, WS) column was prepared. Resin was washed with 12 ml of dHOH and the wash removed after centrifugation at 300 g. The column was then charged with nickel by suspending the resin for 30 minutes in 20 ml of 50 mM $NiSO_4$ and then removing the charging solution by centrifugation. Finally the column was washed with 12 ml of 5 mM imidazole, 500 mM NaCl, 20 mM Tris pH 7.9. Lysate and resin were

then mixed and placed on a shaker which gently agitated (50 rpm) the mix at 4C for 1 hour. The mixture was then centrifuged and the supernatant removed for analysis. The resin and bound his-tagged prostromelysin was then washed five times for 10 minutes with 10 ml of 60 mM imidazole, 500 mM NaCl, 20 mM Tris pH 7.9 at 4C on the shaker. The final mixture was transferred to a poly-prep chromatography column (BioRad, Hercules, CA) and the final wash collected as flow through. Prostromelysin was eluted using 14 ml of 1 M imidizole, 500 mM NaCl, 20 mM Tris pH 7.9. 1 ml fractions were stored at 4C for analysis on 12% polyacrylamide-1% SDS gels. The bulk of the prostromelysin eluted in the second two elution fractions (Figure 24).

To determine the yield of protein from the column, aliquots from elution fractions were analyzed on an electrophoresis gel alongside of a series of lanes with different amounts of BSA. By estimating which BSA lane had a band of equal intensity of the prostromelysin bands we could estimate the amount of protein in the prostromelysin fractions using much less material than a traditional protein assay. From this analysis (Figure 25) we estimate that we recovered 2 BSA equivalent milligrams of prostromelysin per liter of culture.

External Source for Stromelysin. After considerable time and effort in making our own stromelysin by activating prostromelysin and getting fairly low yields we decided to invest in attempts to make the mature form of stromelysin directly from a clone. We constructed expression vector clones containing only the catalytic domain of stromelysin. Thirteen of fourteen isolates tested made active stromelysin. During this period we had been communicating with Dr. Quezang Ye at Parke-Davis about his work with catalytic domain clones and purification methods for stromelysin. We were also interested in getting their clone for stromelysin. After some haggling by the lawyers, Parke-Davis agreed to provide us with stromelysin for the remainder of the project instead of the clone. We were subsequently sent 6.7 mg of stromelysin.

V. Characterization of the scaffold protein (eglin c)

As part of our effort to convert eglin c into a useful inhibitor of other proteins we have carried out a series of investigations into the properties of eglin c as a protein. Those investigations have yielded a method for the quantitative assessment of hypotheses concerning the determinants of protein structure. This method is an extension of the use of mutagenesis and combinatorial libraries to study proteins and we believe the new method has broad utility. We think that this is an important new approach for the study of proteins and are very excited about this development. We anticipate that it will be the focus of research in my laboratory for the next decade. The method is described in the accompanying manuscript which has been submitted to the Proceedings of the National Academy of Science USA (August 30, 1999). The work in this manuscript and the two others that have come from this project are summarized in the sections below.

A Method for the Quantitative Assessment of Hypotheses Concerning the Determinants of Protein Structure (manuscript 3, submitted). Suppose that one had a hypothesis concerning the determinants of some aspect of structure in a particular protein. An approach in use since our discovery of a method for making directed mutations in DNA is to mutagenize the residues involved and determine what happens to the protein. A later adaptation of this approach is to randomize the residues involved in a combinatorial library and assess the consequences both in terms of the fraction of variants in the library that pass whatever tests are applied and in terms of characterization of variants that pass and fail the tests. A more hypothesis oriented approach was

developed in the laboratory of Dr. Michael Hecht in which he constructed combinatorial libraries in which each library member conformed to a hypothesis about the hydrophobic or hydrophilic nature of the residue in a particular position in the protein. Each library variant had the same hydrophilic-hydrophobic pattern at the test sites in the protein. The large fraction of the library variants that passed the tests was then used to assert that the pattern was sufficient to encode structure and that the specific details of the residues were not critical.

We have extended this approach by the addition of two features; one modifies the library construction approach and the other changes the mode of analysis of the library. First, we use resin-splitting technology (29,30) to facilitate the construction of arbitrarily complex libraries that are free of the constraints imposed by the genetic code. In all of the previous studies the libraries that were made were all rendered degenerate either by randomizing residues or by using degenerate codons (codons with mixed nucleotides at one or more sites). This limits the nature of the libraries that can be constructed. Split-resin technology can be used to synthesize an arbitrarily complex set of degenerate oligonucleotides and frees library construction from the previous constraints.

The second feature that we have added to the assessment of libraries involves the use of regression analysis to extend the analytical power of combinatorial library experiments. Appropriate selection of the nature of variants in such a library makes it possible to use regression analysis for quantitative assessment of specific hypotheses and, by averaging over the effects of many factors, to extract accurate information regarding partial effects contributing to protein structure formation. Regression analysis can also be used to assess several competing hypotheses using a single library, in contrast to the approach using the fraction of the library variants that remains active as the metric. Regression analysis provides access to new information by providing a formalism for the quantitative evaluation of the consequences of the effects defined in a hypothesis and a statistical assessment of the degree to which variant behavior can be attributed to them.

With this approach we have shown that we can independently assess and reproduce a known determinant of protein structure, α-helix propensities. Helix propensities represent the contribution to the free energy of denaturation made by the various amino acids in solvent exposed positions in an α-helix. Regression parameters derived from the analysis of a 455 member sample from a library wherein four solvent-exposed sites in an α-helix can contain any of nine different amino acids are highly correlated ($P < 0.0001$, $R^2 \geq 0.97$) to the relative helix propensities for those amino acids, as estimated by a variety of biophysical and computational techniques. This agreement encourages us to believe that our approach can provide quantitative assessments of other hypotheses about the relations between amino acid sequence, stability and structure.

Effect of a Poly-His-Terminal Extension on Eglin c Stability (manuscript #1; Anal. Biochem 1998). To facilitate protein purification, a poly-his-terminal extension has been incorporated onto eglin c. In this manuscript we reported that the his-tag incorporation does not effect eglin c stability. Thermal denaturations monitored by circular dichroism spectropolarimetry showed that the free energy of denaturation did not change upon his-tag incorporation.

Nonideality and Protein Thermal Denaturation (manuscript #2; Biopolymers 1999). We studied the thermal denaturation of eglin c by suing CD spectropolarimetry and differential scanning calorimetry (DSC). At low protein concentrations, denaturation is consistent with the classical two-state model. At

concentrations greater than several hundred uM, however, the calorimetric enthalpy and the midpoint transition temperature increase with increasing protein concentration. These observations suggested the presence of intermediates and/or native state aggregation. However, the transitions are symmetric, suggesting that intermediates are absent, the DSC data do not fit models that include aggregation, and analytical ultracentrifugation (AUC) data show that native eglin c is monomeric. Instead, the AUC data show that eglin c solutions are nonideal. Analysis of the data gives a second virial coefficient that is close to values calculated from theory and the DSC data are consistent with the behavior expected from nonideal solutions. We conclude that the concentration dependence is caused by differential nonideality of the native and denatured states. This nonideality is hypothesized to arise from the high charge of the protein at acid pH and is exacerbated by the low buffer conditions in which these experiments are traditionally carried out.. Our conclusions may explain differences between van't Hoff and calorimetric denaturation enthalpies observed for other proteins whose behavior is otherwise consistent with the classical two-state model.

KEY RESEARCH ACCOMPLISHMENTS

- we have discovered that eglin c is not a good scaffold protein for phage display
- we have discovered that a circularly permuted version of eglin (peglin) also does not function well as a scaffold protein in phage display
- we have discovered a new way make and analyze combinatorial libraries to assess hypotheses concerning the determinants of protein structure

REPORTABLE OUTCOMES

Manuscripts

Published

1. Waldner, JC, Lahr, SJ, Edgell, MH, and Pielak, GJ. Effect of a polyhistidine terminal extension on eglin c stability. *Anal. Biochem.* **263** 116-118 (1998)
2. Waldner, JC, Lahr, SJ, Edgell, MH, and Pielak, GJ. Nonideality and protein thermal denaturation. *Biopolymers* **49** 471-479 (1999)

Submitted

3. Lahr, SJ, Broadwater, A, Carter, CW, Jr., Collier, ML, Hensley, L, Waldner, JC, Pielak, GJ, and Edgell, MH. Patterned library analysis: a method for the quantitative assessment of hypotheses concerning the determinants of protein structure. *Proc. Nat. Acads.Sci. USA* (submitted 1999)

Patents

none

Degrees

Mr. Steven Lahr is writing his PhD thesis. The research that will be reported in that thesis is the biophysical characterization of eglin c and the development of our method for analyzing combinatorial libraries which we are calling patterned library analysis.

Cell lines

none

Informatics

none

Funding Applied for

I applied to NIH in 1998 for funding to explore and apply the patterned library analysis method. That proposal was not funded but the review indicated that they liked the approach but felt the method needed the assessment of publication. I intend to submit an amended proposal for the November 1 NIH deadline.

Employment opportunities

Mr. Lahr has secured a post-doctoral position with Dr. William DeGrado and will join that laboratory in October 1999.

CONCLUSIONS

Although we were not able to develop an appropriate scaffold protein that could be used to generate high affinity inhibitors for proteins involved in the metastasis of cancer I feel that the general concept for making inhibitors is still viable. This is based on the fact that we were able to accomplish all of the parts of the project except for the very crucial part of developing an appropriate scaffold protein. A different scaffold protein or a more extensively engineered version of eglin might be expected to work. We had anticipated that the difficult part of the project would be to recover weak binders from the target proteinase since weak binders might be expected to be substrates, to be cleaved and hence to not bind. However, we were able to find binders to stromelysin or papain in the first pass of panning with a phage display library containing a site cleavable by the target. What else have we learned about making protein based inhibitors and stromelysin?

1. Prostromelysin can be made in large quantities in *E. coli*. Most of the product is in a soluble form, that is, not inclusion bodies. The protein is not processed to a toxic form in the bacterial cell.

2. *In vitro* activation of prostromelysin with trypsin is very inefficient. The best we could ever do was convert about 10% of prostromelysin to catalytically active stromelysin.

3. The catalytic portion of the stromelysin protein when produced in *E. coli* is active. That is, the protein does not need the pro-sequence to fold correctly.

4. We were able to find peptide binders to more than half of the 'randomly' chosen enzymes in our test panel. If those binding sites are in the active site of those enzymes then it should be possible to construct protein based inhibitors for a very wide range of enzymes.

What else have we learned? As part of our efforts to characterize the binding epitope display framework protein, eglin c, we discovered a new approach to assess hypotheses concerning the determinants of protein structure. This approach is based on two realizations. The first is that one can build combinatorial libraries encoding proteins that are free of the constraints of the genetic code by using an existing technology called resin splitting to synthesize the desired degenerate oligonucleotides. The second is that an increased level of quantitative information could be extracted from combinatorial libraries using regression analysis. Employing these insights we carried out a proof-of-principle experiment using the new method in which we were able to reproduce known values for α-helix propensity indicating that the method does indeed work. Mutagenesis is the major tool for exploring hypotheses concerning the determinants of protein structure. Any advance in our capacity to use this tool should have broad utility. We are very hopeful that the approach that we have discovered is such an advance.

PERSONNEL SUPPORTED (In full or in part)

Anne Broadwater	(Technician)
Martha L. Collier	(Technician)
Dr. Marshall Hall Edgell	(Principal investigator)
Lucinda Hensley	(Technician)
Steven J. Lahr	(Graudate student)
Paula Davis-Searles	(Graduate student)
Dr. Gary Pielak	(co-Principal investigator)
Jennifer C. Waldner	(Graduate student)

PERSONNEL WHO WORKED ON THE PROJECT WITHOUT SUPPORT

Jill Sherman	(Undergraduate student)
Sharon Smith	(Master's student, Carolina Central University)
Amy Turner	(Undergraduate student)

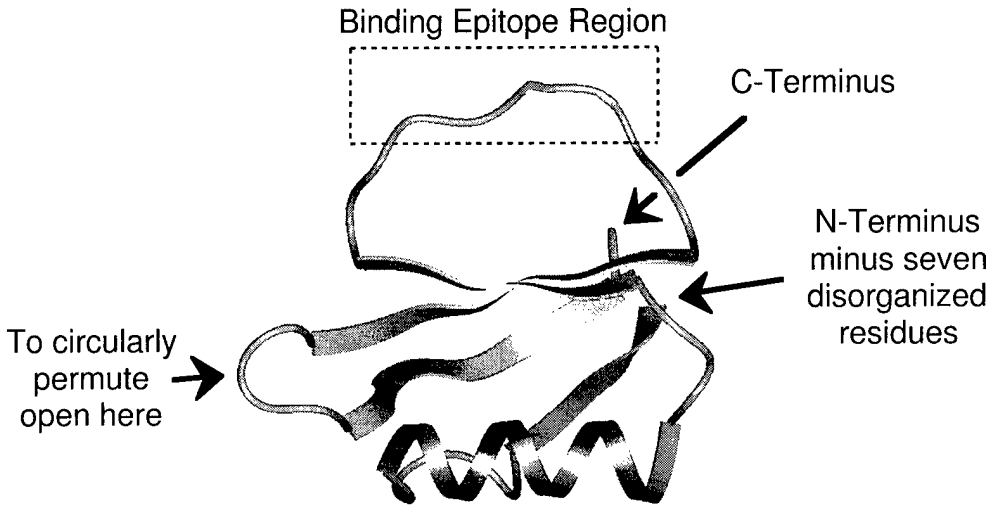


FIGURE 1. Ribbon Diagram of Eglin C. Note that attaching a phage particle to the C-terminus of eglin c might block access to the binding epitope. Our construction of a circularly permuted eglin removes the seven disorganized residues from the N-terminus, adds a four residue tight turn to connect the N- and C-terminal ends, and opens the protein to create new N- and C-termini at the point indicated in the figure.

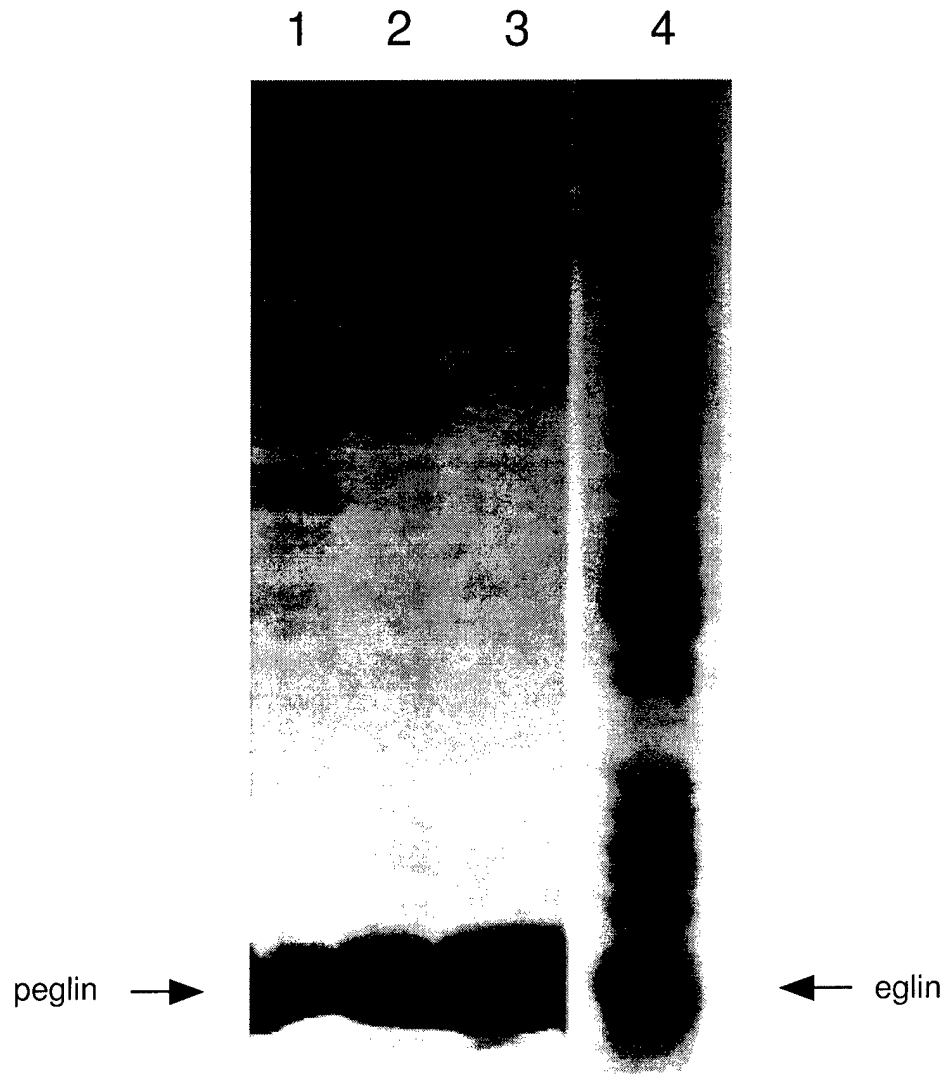


Figure 2. *E coli* lysates expressing pegin or eglin c.

Lane 1. Extract from BLR with a gene for PEGLIN 0 hr. after induction
Lane 2. Extract from BLR with a gene for PEGLIN 1.5 hr. after induction
Lane 3. Extract from BLR with a gene for PEGLIN 3.0 hr. after induction
Lane 4. Extract from BLR with a gene for EGLIN C 3.0 hr. after induction

15% Polyacrylamide gel with 1% SDS.

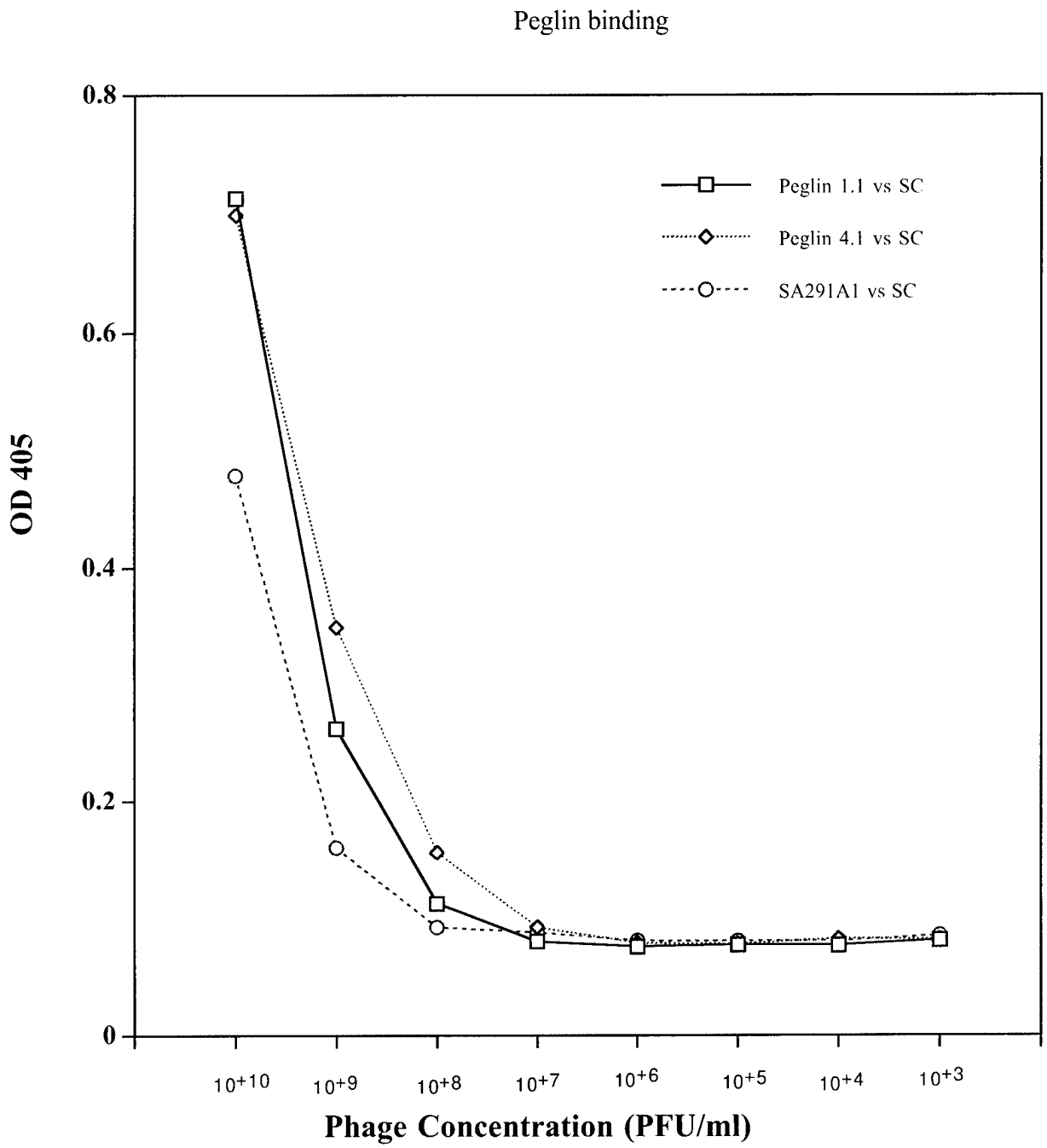


Figure 3. Binding to Wells Coated with Subtilisin. The two peglin constructs both bind to the native target for eglin c, subtilisin, more readily than does a phage displaying a peptide which binds to streptavidin. Since all proteins bind to some extent to this proteinase we wanted to determine relative binding efficencies. Binding at high phage concentrations is due to non-specific binding.

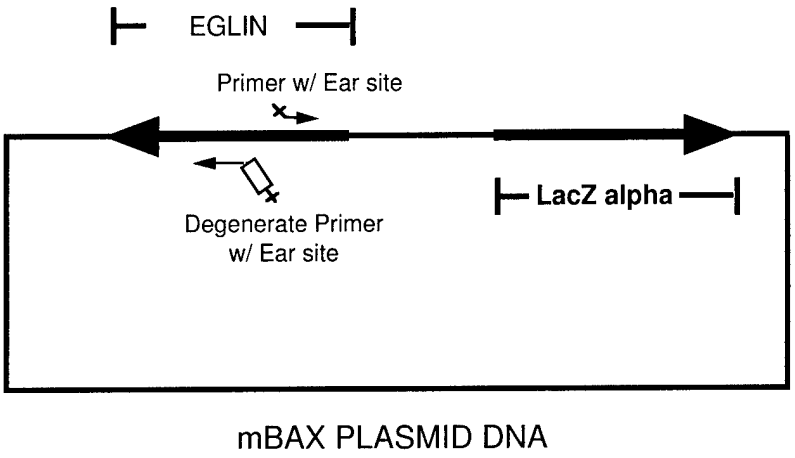


Figure 4. EarI strategy for library construction. PCR can be used to amplify an entire plasmid DNA. If one of the primers is degenerate (the unfilled box represents the degenerate region) then the amplified products will contain that degeneracy. By using sites on the ends of the primers (bars show sites) such as EarI or EamI that cut outside of their recognition sites one can produce amplified product that when cleaved and ligated contains no restriction site sequence. That is, one can cut and ligate the product DNA at chosen sites that are totally independent of the distribution of sites in the template DNA as long as there are not additional EarI sites in the template.

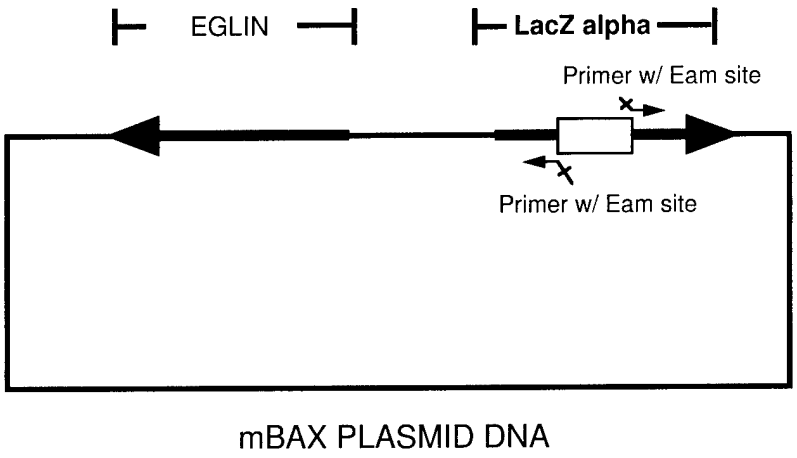


Figure 5. Construction of a 'White' mBAX Phage. A ~200 bp deletion in the alpha fragment of beta-galactosidase within the mBAX plasmid was created by using PCR and primers cotaining Eam restriction sites. Restriction enzymes recognizing these sites cut outside of the recognition site and hence allow one to cut and splice anywhere one likes within a sequence indepedent of the distribution of restriction sites in the target DNA. This presumes that there are not additional Eam sites in the target DNA. Circular plasmid DNA was used as the template for PCR.

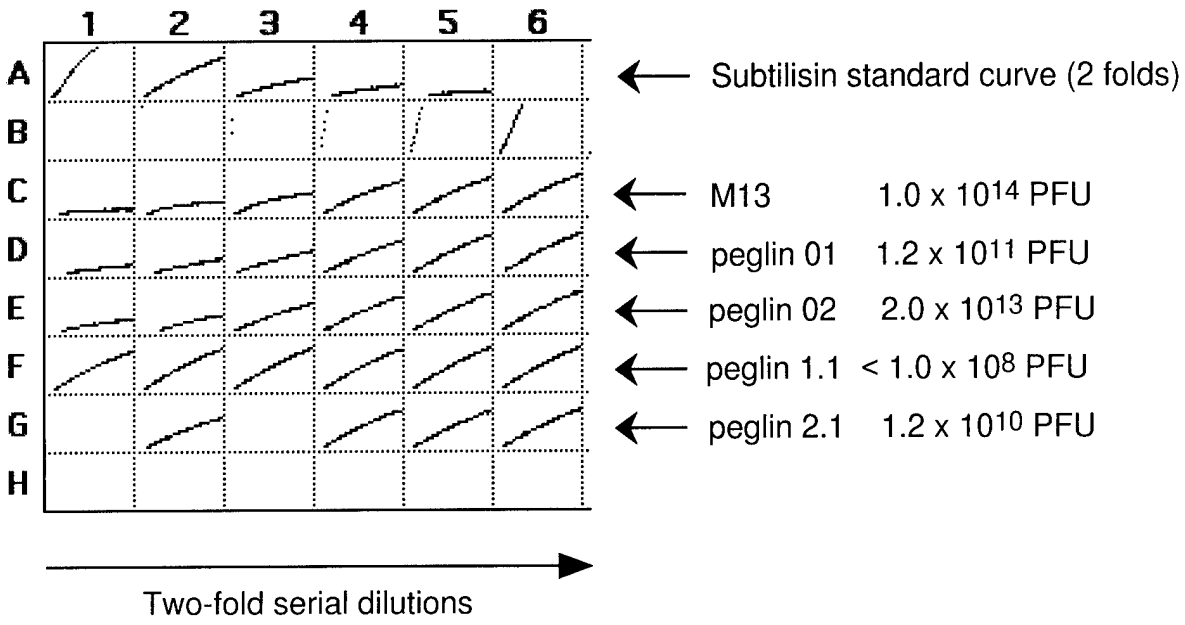


Figure 6. Inhibitory Effects of Various Phage preparations without Dilution (i.e. similiar amounts of PEG). Aliquots of PEG concentrated phage stocks were mixed with 8.7 x 10^8 molecules of subtilisin for 60 minutes. Substrate was then added and cleavage of substrate followed by color development at 405 nm. Reduced color development indicates inhibition. The fact that each phage preparation shows essentially the same amount of inhibition (slopes in column 2 or 3) independent of the number of phage inthe well suggests that the inhibition is not due to phage.

Wells G1 and G3 artifactually show no color development (presumably no substrate got added).

MOLECULAR DEVICES
Raw Data (Plate)

DATA FILE: DATA 9/24 11_14_47
DESCRIPTION: 9/24/98mlcPeglinclones1.1-5.2vsSubtlisin&catalase plateA
PROTOCOL: PRINTED: 9/24/98
DESCRIPTION: Phage: 18 hr ON -30 min color development
MODE: Endpoint AUTOMIX: ON
WAVELENGTH: 405 CALIBRATION: ON
MEAN TEMP: 24.90°C SET TEMP: OFF

Optical Density

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.064	0.068	0.074	0.074	0.092	0.081	0.103	0.078	0.133	0.080	0.126	0.088
B	0.068	0.067	0.072	0.072	0.087	0.072	0.102	0.085	0.118	0.081	0.138	0.091
C	0.064	0.060	0.071	0.072	0.085	0.069	0.085	0.068	0.107	0.075	0.121	0.078
D	0.065	0.062	0.074	0.069	0.085	0.083	0.087	0.068	0.111	0.072	0.113	0.090
E	0.168	0.137	0.125	0.133	0.133	0.127	0.140	0.124	0.190	0.136	0.200	0.161
F	0.100	0.118	0.110	0.114	0.127	0.125	0.137	0.121	0.173	0.135	0.219	0.143
G	0.085	0.091	0.092	0.099	0.114	0.103	0.116	0.105	0.156	0.110	0.165	0.113
H	0.092	0.090	0.090	0.097	0.107	0.098	0.113	0.098	0.135	0.102	0.158	0.110

Figure 7a. Binding (or lack of binding) of Peglin to Stromelysin. Rows A through D are wells coated with subtilisin. Rows E through H are coated with catalase to which the peglin clones should not bind. Odd columns have 1:5 dilutions of a phage stock and even rows have 1:50 dilutions. The different dilutions allow one to assess phage concentration dependent binding.

Columns 1,2 contain peglin isolate v1.1
Columns 3,4 isolate v1.2
Columns 5,6 isolate v2.1
Columns 7,8 isolate v2.2
Columns 9,10 isolate v3.1
Columns 11,12 isolate v3.2;

A phage preparation containing phage that bound to its cognate target (subtilisin) would have larger numbers in rows A-D than rows D-H.

MOLECULAR DEVICES
Raw Data (Plate)

DATA FILE: DATA 9/24 11_10_45
DESCRIPTION: 9/24/98mlcPeglinclones1.1-5.2vsSubtlisin&catalase-plateB
PROTOCOL: PRINTED: 9/24/98
DESCRIPTION: Phage: 18 hr ON -30 min color development
MODE: Endpoint AUTOMIX: ON
WAVELENGTH: 405 CALIBRATION: ON
MEAN TEMP: 24.80°C SET TEMP: OFF

Optical Density

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.064	0.062	0.075	0.067	0.081	0.063	0.066	0.062	0.059	0.085	0.356	0.078
B	0.060	0.054	0.077	0.059	0.075	0.060	0.068	0.061	0.061	0.057	0.285	0.081
C	0.063	0.056	0.093	0.058	0.074	0.058	0.063	0.063	0.063	0.060	0.301	0.084
D	0.069	0.058	0.082	0.062	0.082	0.063	0.069	0.064	0.066	0.071	0.367	0.095
E	0.157	0.106	0.359	0.183	0.131	0.121	0.111	0.110	0.149	0.123	0.482	0.208
F	0.152	0.100	0.369	0.192	0.126	0.111	0.120	0.112	0.150	0.122	0.489	0.248
G	0.177	0.109	0.373	0.195	0.141	0.119	0.121	0.112	0.152	0.128	1.909	2.015
H	0.178	0.115	0.412	0.198	0.141	0.124	0.119	0.117	0.161	0.151	1.635	2.019

Figure 7b. Binding (or lack of binding) of Peglin to Stromelysin. Rows A through D are wells coated with subtilisin. Rows E through H are coated with catalase to which the peglin clones should not bind. Odd columns have 1:5 dilutions of a phage stock and even rows have 1:50 dilutions. The different dilutions allow one to assess phage concentration dependent binding.

Columns 1,2 contain peglin isolate v4.1
Columns 3,4 isolate v4.2
Columns 5,6 isolate v5.1
Columns 7,8 isolate v5.2;

Wells G11,12 and H11,12 were loaded with a catalase binding phage to serve as our phage binding controls.

A phage preparation containing phage that bound to its cognate target (subtilisin) would have larger numbers in rows A-D than rows D-H.

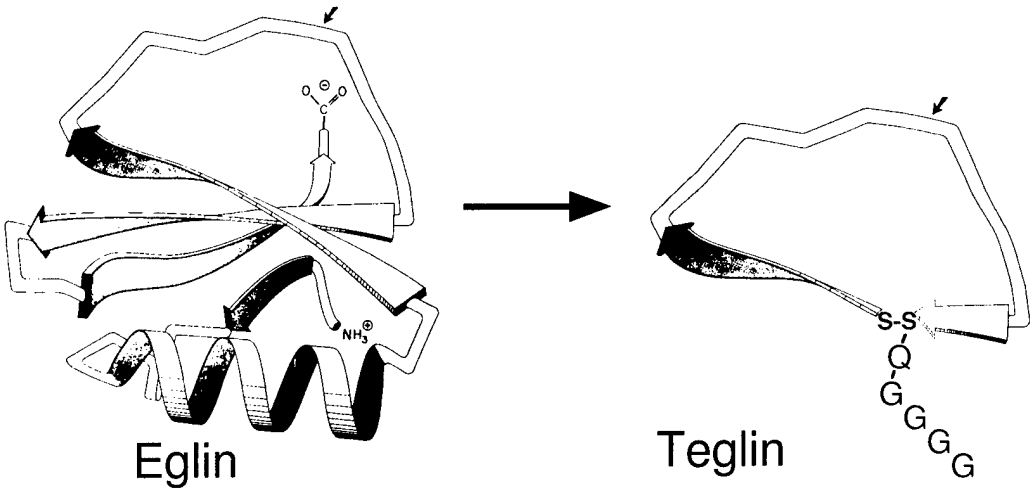


Figure 8. Conversion of Egin to Teglin. The ten amino acid loop containing the binding epitope and the underlying beta strands containing residues that contribute to two framework to loop salt-bridges is removed and joined via a cysteine bond. The teglin structure would not of course maintain the beta sheet conformation. The closed loop structure is fused to the M13 pIII protein via a QGGGG linker peptide.

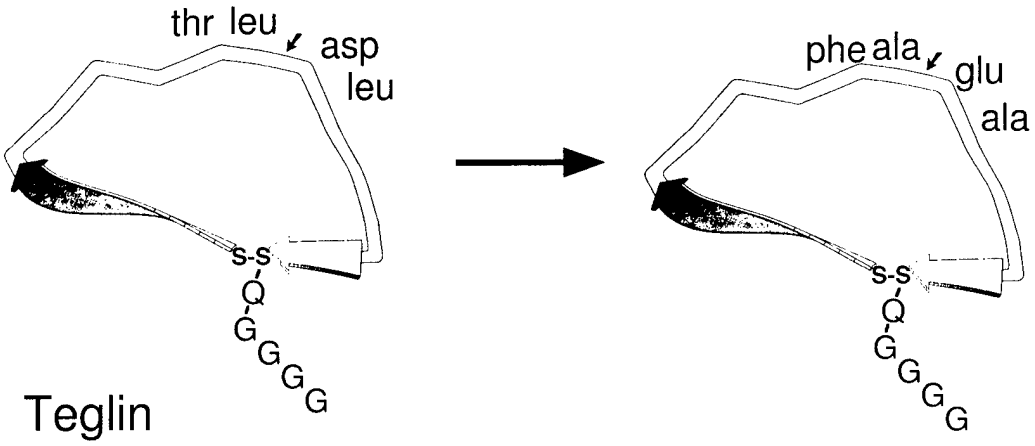


Figure 9. Conversion of Tegin to Teglin-papain. The two resdiues in teglin on each side of what would be the cleavage site in a substrate (indicated by the arrow) were replaced in teglin-papain with the residues in a papain cleavage site.

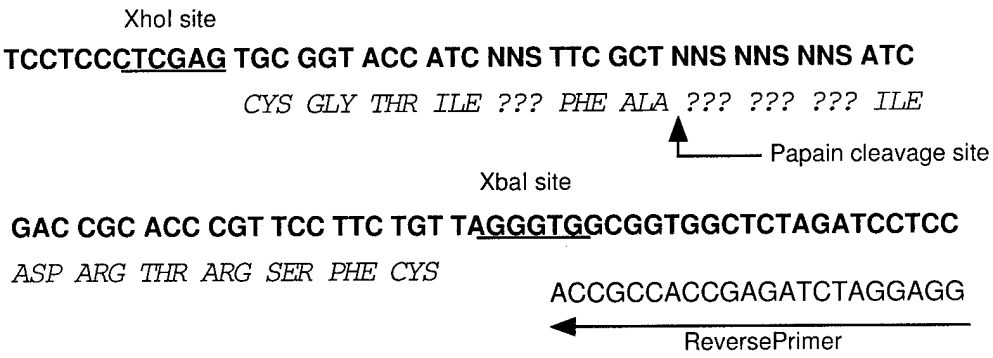


Figure 10. Construction of a Teglin Based Papain Cleavge Sequence Library. An ologonucleotide was syntehsized with three randomized codons just downstream of the papain binding site and one randomized codon just upstream of the cleavage site. The oligonucleotide was rendered double stranded using the primer shown. The double stranded DNA was then cleaved with Xho and Xba and inserted into the M13 display phage containing the teglin gene such that the native subtilisin binding eptiope is replaced by the degenerate sequence.

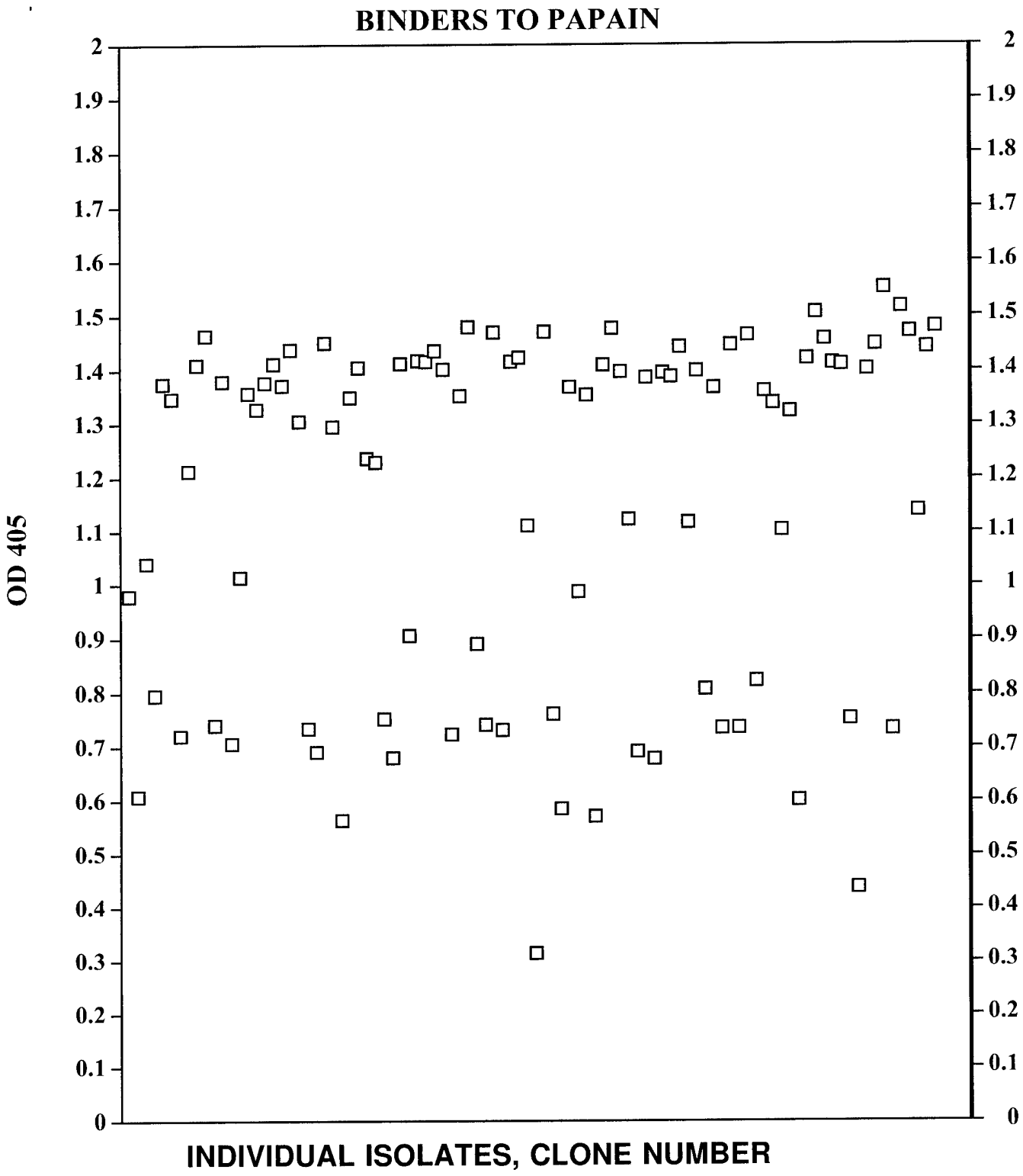


Figure 11. ELISA Values for Individual Clones After Six Rounds of Panning Against Papain. Individual clones are isolated at random from the population which was present after six rounds of panning against papain and bound to a well in a 96-well plate coated with papain. The amount of phage that binds is determined by an ELISA assay using a primary antibody against wild-type M13 bacteriophage.

Design

TGC GGT ACC ATC NNS TTC GCT NNS NNS NNS ATC GAC CGC ACC CGT TCC TTC TGT
cys gly thr ile xxx phe ala xxx xxx xxx ile asp arg thr arg ser phe **cys**

Higher Affinity Papain Binders

TGG GTA CCA TCG GGT TCG CTG GGA CGC GGG ATC GAC CGC ACT TGT TCC TTC TGT
trp val pro ser gly ser leu gly arg gly ile asp arg thr cys ser phe **cys**

TGG CGG TAC CAT CAA GTT CGC TCG GAG GCT GAT CGA CCG CAC CCG TTC CTT CTT
trp arg tyr his gln val arg ser glu ala asp arg pro his pro phe leu leu

TGC GGT ACC ATC GGG TTC GCT CCG AGG CTG ATC GAC CGC ACC CAT TCC TTC TTT
cys gly thr ile gly phe ala pro arg leu ile asp arg thr his ser phe phe

TGG CGG TAC CAT CAC GTT CGC TCC GAG CCC GAT CGA CCG CAC CCG TTC CTT CTG
trp arg tyr his his val arg ser glu pro asp arg pro his pro phe leu leu

TGC GGT ACC ATC GAC TTC GCT AAG AGG ACG ATC TAC CGC ACC CAT TCC TTC TGG
cys gly thr ile asp phe ala lys arg thr ile tyr arg thr his ser phe trp

Intermediate Affinity Papain Binders

TAC CAT CTA GTT CGC TGG GGG GAG GAT CGA CCG CAC CCG TTC CTT CTG TCG GGT
tyr his leu val arg trp gly glu asp arg pro his pro phe leu ser gly gly

TGC GGT ACC ATC GAG TTC GCT GGG GGC GGG ATC GAC CGC ACC CGT TCC TTC TGT
cys gly thr ile glu phe ala gly gly gly ile asp arg thr arg ser phe **cys**

TGC GGT ACC ATC TGG TTC GCT GGG GGG GAT ATC GAC CGC ACC CGT TCC TTC TGT
cys gly thr ile trp phe ala gly gly asp ile asp arg thr arg ser phe **cys**

Figure 12. Sequences of Binders to Papain. All but two of the sequences have undergone some mutation relative to the library design which has removed the possbility of forming a constrained loop through a disulphide bond.

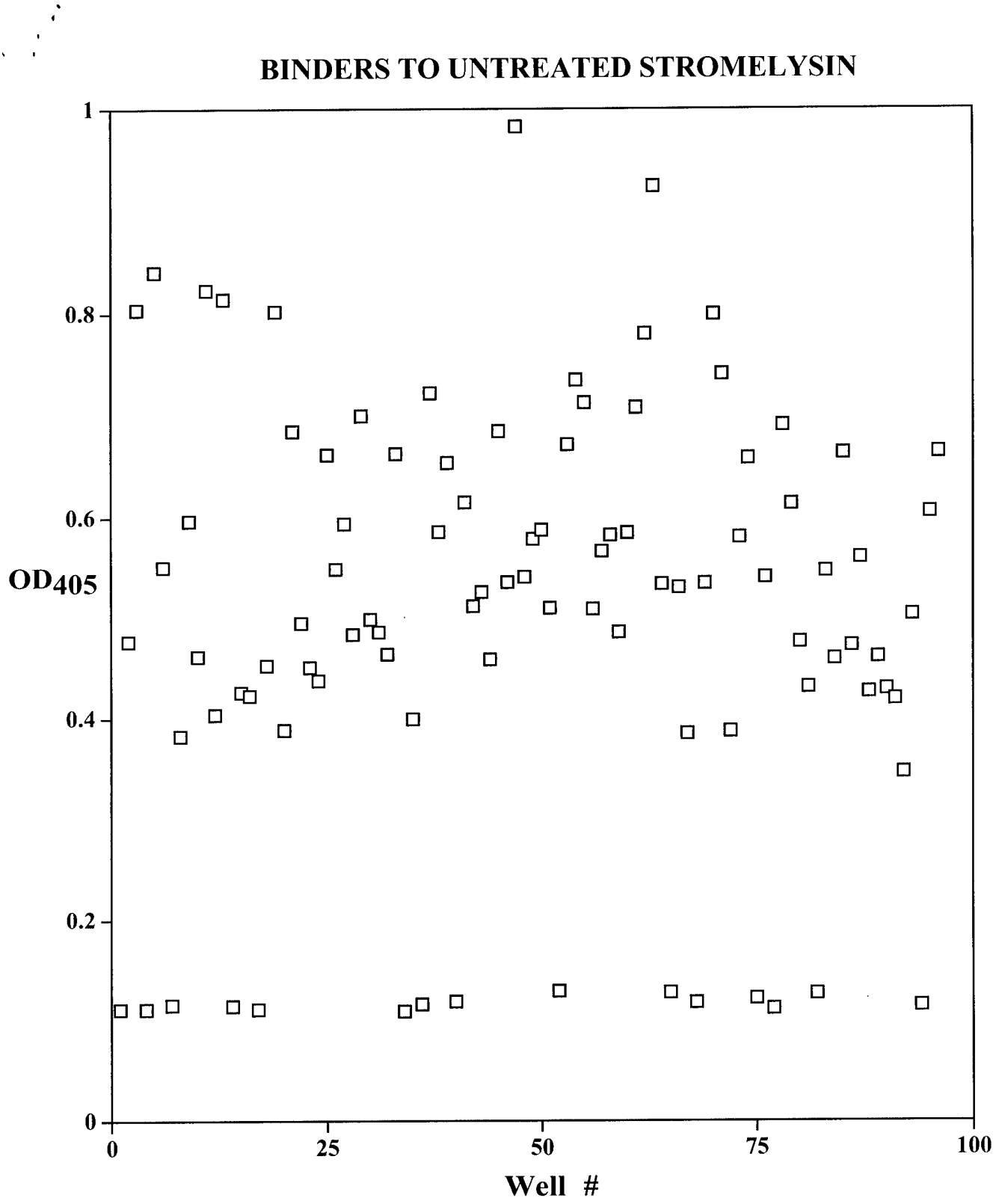


Figure 13. ELISA Values for Individual Clones After Six Rounds of Panning Against Stromelysin. Individual clones are isolated at random from the population which was present after six rounds of panning against stromelysin and bound to a well in a 96-well plate coated with stromelysin. The amount of phage that binds is determined by an ELISA assay using a primary antibody against wild-type M13 bacteriophage.

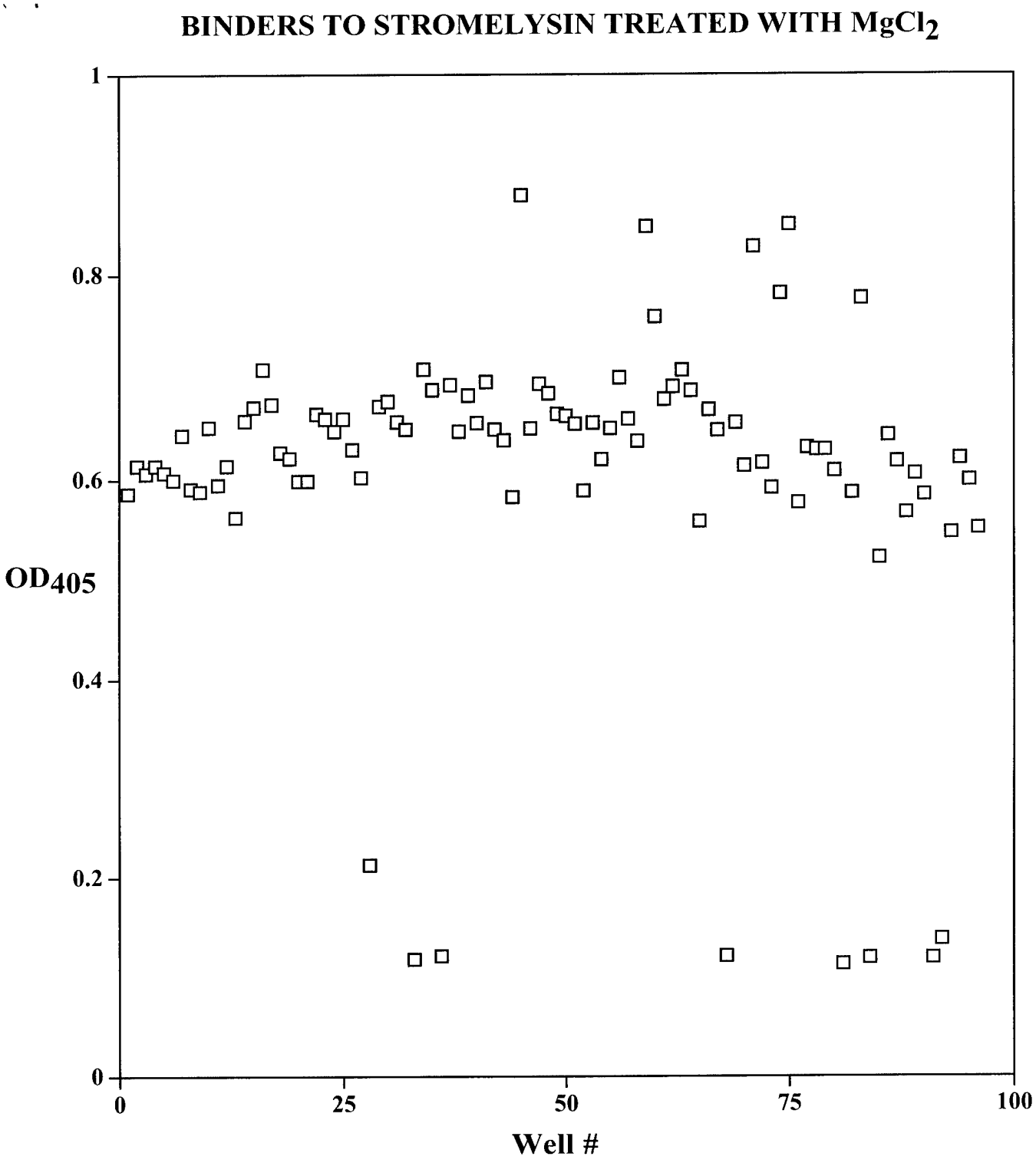


Figure 14. ELISA Values for Individual Clones After Six Rounds of Panning Against Stromelysin Treated with $MgCl_2$. Individual clones are isolated at random from the population which was present after six rounds of panning against $MgCl_2$ treated stromelysin and bound to a well in a 96-well plate coated with stromelysin. The amount of phage that binds is determined by an ELISA assay using a primary antibody against wild-type M13 bacteriophage.

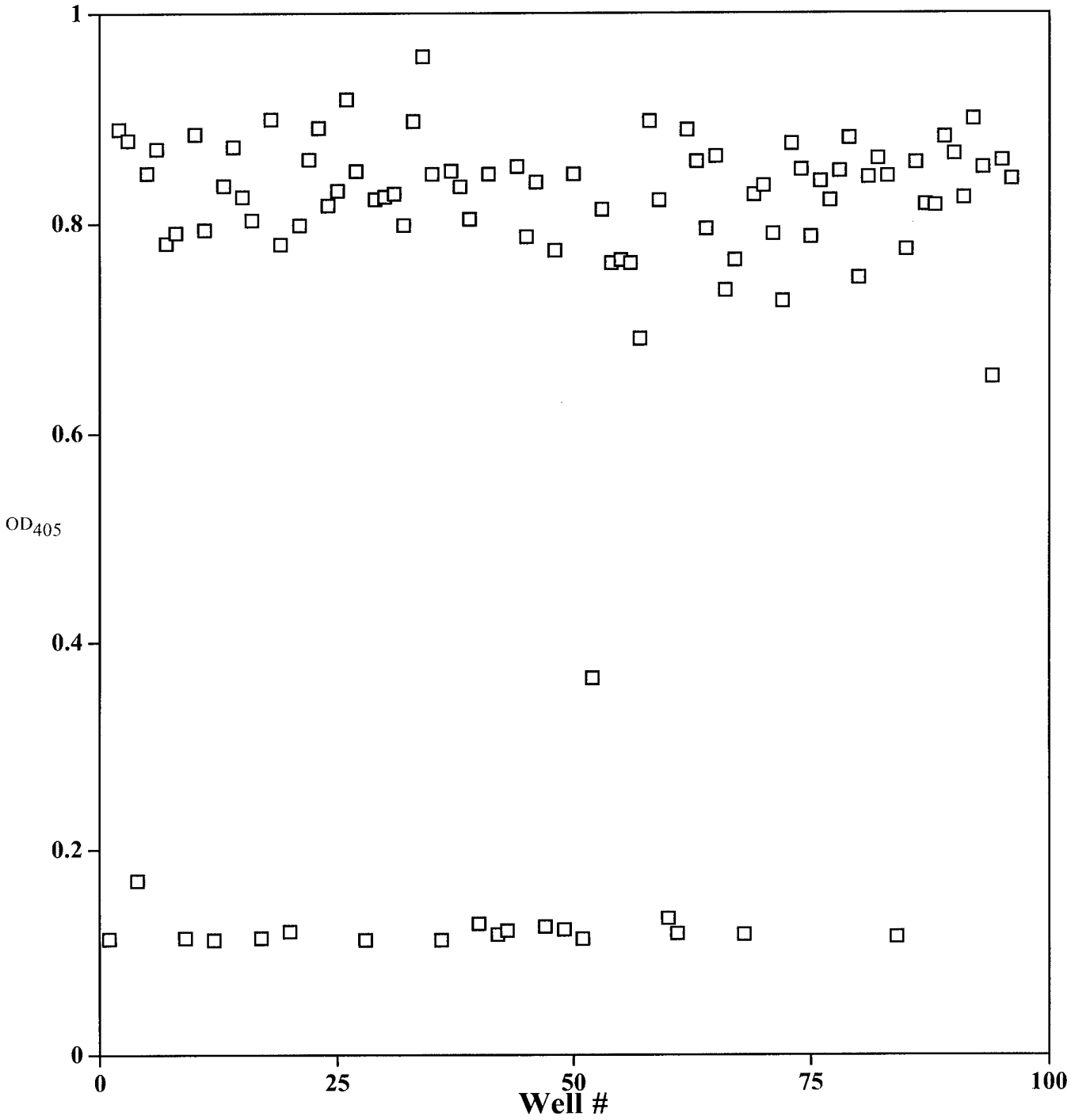


Figure 15. ELISA Values for Individual Clones After Six Rounds of Panning Against Stromelysin Treated with $CdCl_2$. Individual clones are isolated at random from the population which was present after six rounds of panning against $CdCl_2$ treated stromelysin and bound to a well in a 96-well plate coated with stromelysin. The amount of phage that binds is determined by an ELISA assay using a primary antibody against wild-type M13 bacteriophage.

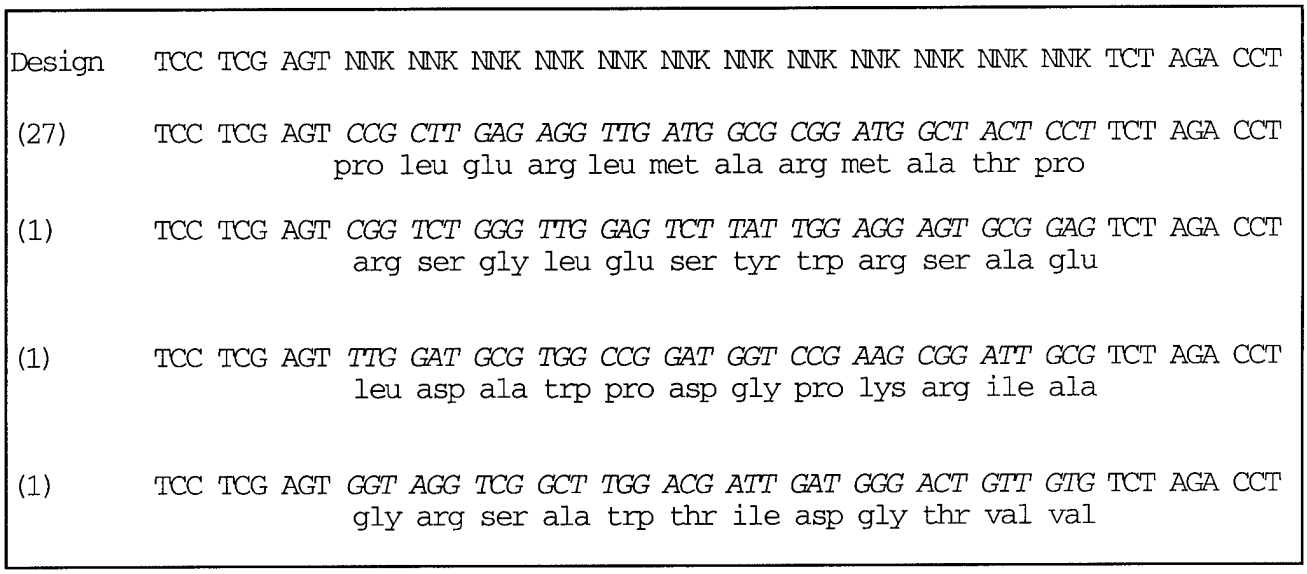


Figure 16. Sequences of Clones Which Bound to $CdCl_2$ Treated Stromelysin. Most of the isolates had a single sequence. The number of isolates with a given sequence are indicated in parenthesis.

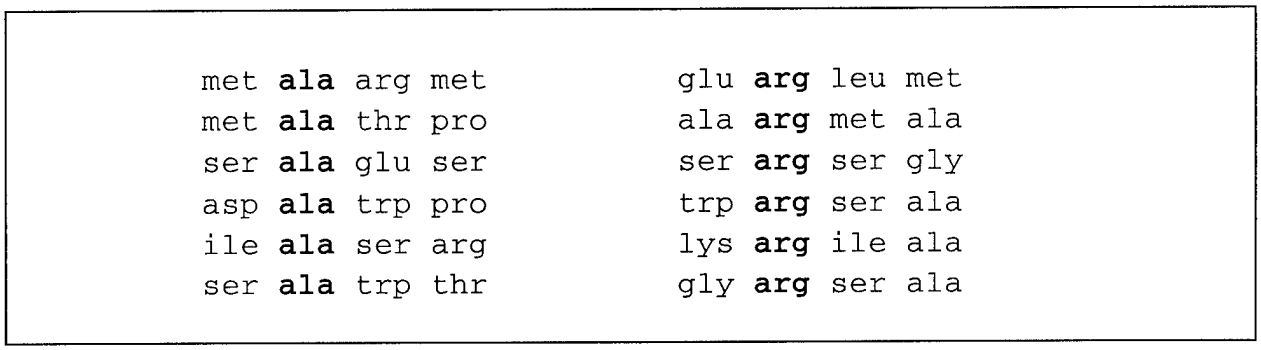


Figure 17. Alignments of Sequences from Stromelysin Binding Clones. Only alanine and arginine are present in the variant portion of all of the isolates. Alignments are presented around each of these two residues.

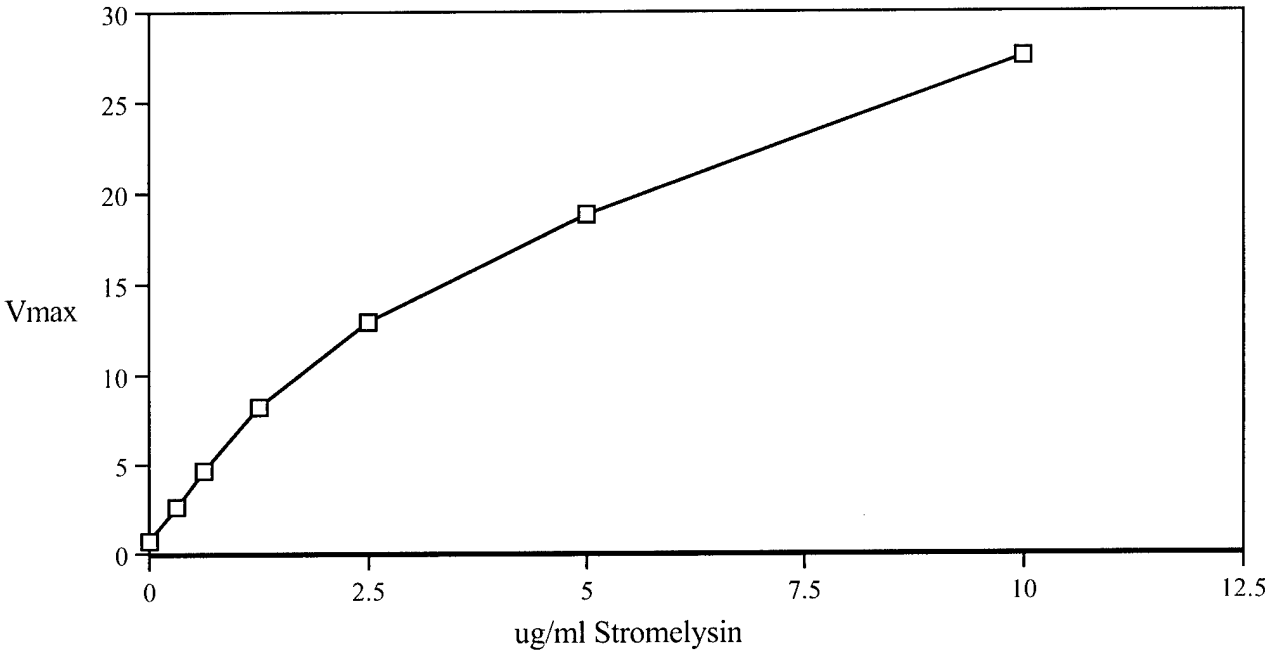


Figure 18. Standard Curve for Stromelysin. A colormetric microtiter plate assay was used to monitor amounts of stromelysin. A kinetic curve of color delvelopment using as substrate, Ac-Pro-Leu-Gly-[2-mercapto-4-methyl-pentanoyl]-Leu-Gly-0Et, was collected in wells of a microtiter plate at 405 nm in a Molecular Devices plate reader. Vmax was calculated and used as the metric for amounts of stromelysin. Reactions were at room temperature for 10 minutes with readings taken every 11 seconds.

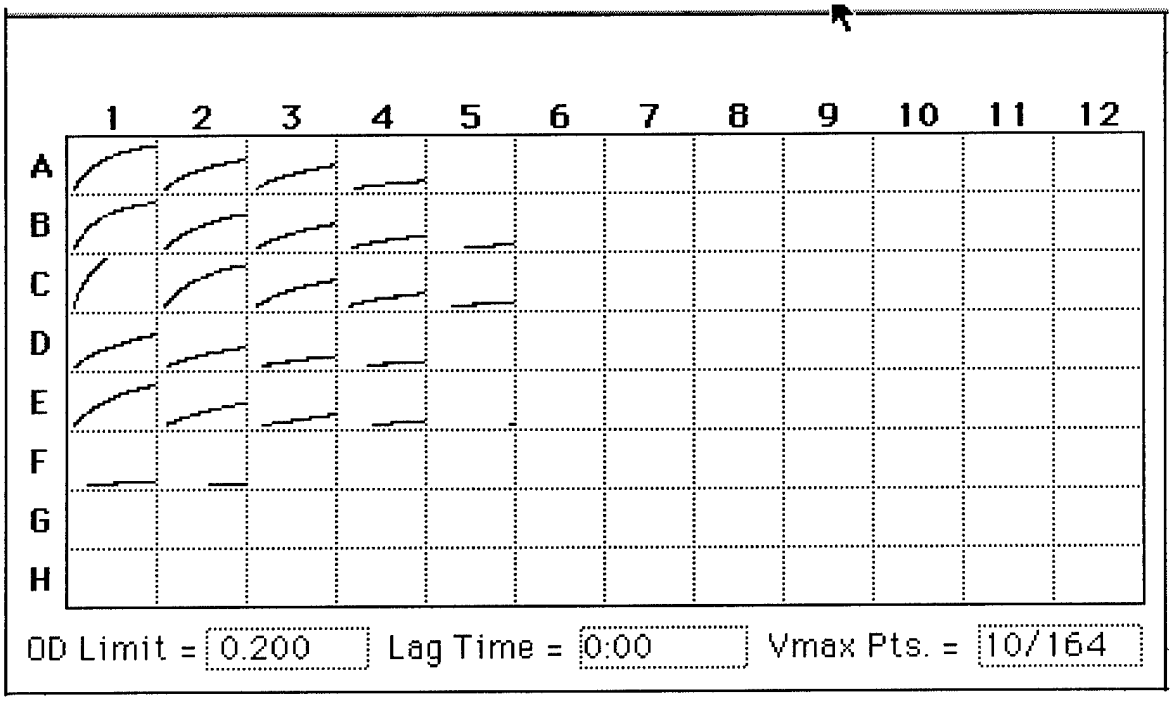
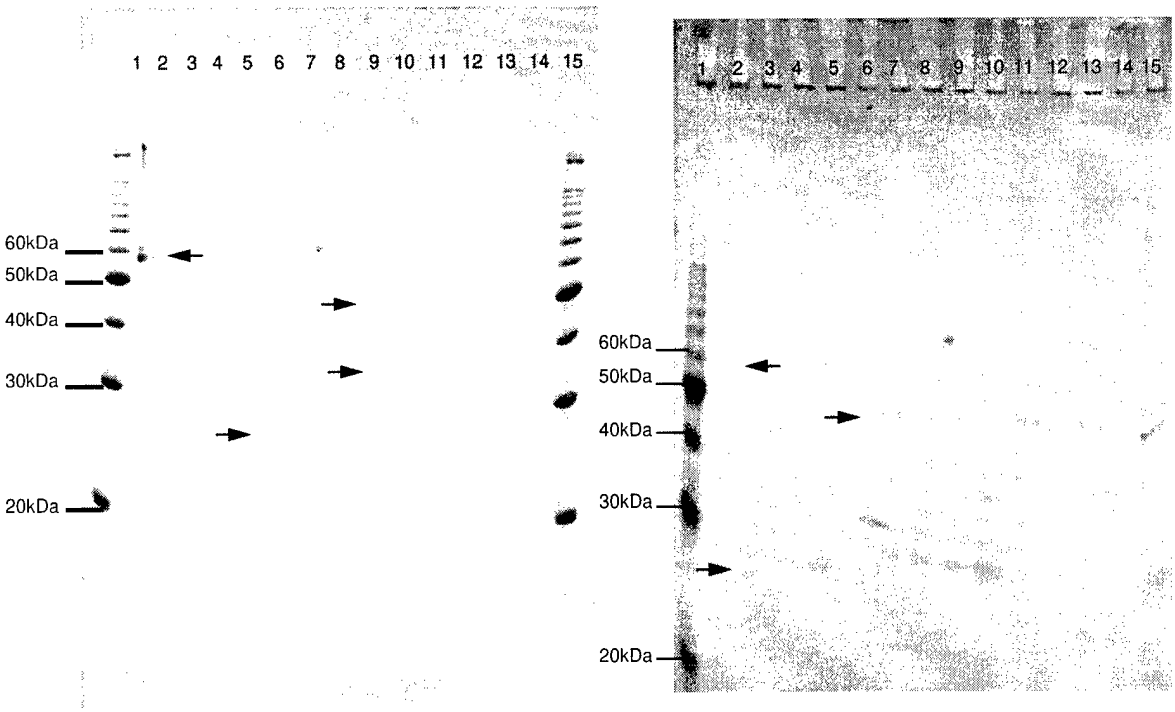


Figure 19. Trypsin Activation of Prostromelysin. Samples of mature stromelysin and trypsin treated prostromelysin assayed with the thiopeptide substrate. Each box shows the color production as a function of time in wells of a microtiter plate. Readings were taken every 11 seconds. Reactions were carried out in 50 mM Tris pH 7.4, 5 mM $CaCl_2$ and 200 mM NaCl. Trypsin cleavage was for 30 minutes at 37C.

Rows A-C Mature Stromelysin
- column 1: 5.0 ug/ml stromelysin
- column 2: 2.5 ug/ml stromelysin
- column 3: 1.2 ug/ml stromelysin
- column 4: 0.6 ug/ml stromelysin

Rows D,E Prostromelysin treated with 25 uM trypsin
- column 1: 100 ug/ml prostromelysin
- column 2: 50 ug/ml prostromelysin
- column 3: 25 ug/ml prostromelysin
- colimn 4 12 ug/ml prostromelysin

Row F Prostromelysin treated with 75 uM trypsin
- column 1: 100 ug/ml prostromelysin
- column 2: 50 ug/ml prostromelysin
- column 3: 25 ug/ml prostromelysin
- colimn 4 12 ug/ml prostromelysin



1 10kDa Marker
2 Prostromelysin
3 Prostrom. 10min. 100μM trypsin
4 Prostrom. 20min. 100μM trypsin
5 Prostrom. 30min. 100μM trypsin
6 Prostrom. 10min. 25μM trypsin
7 Prostrom. 20min. 25μM trypsin
8 Prostrom. 30min. 25μM trypsin
9 Prostrom. 10min. 4.16μM trypsin
10 Prostrom. 20min. 4.16μM trypsin
11 Prostrom. 30min. 4.16μM trypsin
12 Prostrom. 45min. 100μM trypsin
13 Prostrom. 45min. 25μM trypsin
14 Prostrom. 45min. 4.16μM trypsin
15 10kDa Marker

1 10kDa Marker
2 Prostromelysin
3 Prostrom. 10min. 4.16μM trypsin
4 Prostrom. 20min. 4.16μM trypsin
5 Prostrom. 30min. 4.16μM trypsin
6 Prostrom. 10min. 4.16μM trypsin +PMSF
7 Prostrom. 10min. 2μM trypsin
8 Prostrom. 20min. 2μM trypsin
9 Prostrom. 30min. 2μM trypsin
10 Prostrom. 60min. 2μM trypsin
11 Prostrom. 10min. 1μM trypsin
12 Prostrom. 20min. 1μM trypsin
13 Prostrom. 30min. 1μM trypsin
14 Prostrom. 50min. 1μM trypsin
15 Prostrom. 60min. 1μM trypsin

Figure 20. Gel Assay for Trypsin Activation of Prostromelysin. Prostromelysin has a molecular weight of 58 kD. Mature stromelysin is 45 kD and a processed active form at 28 kD. Digestions were done in 50 mM Tris pH 7.4, 5 mM $caCl_2$, 200 mM NaCl,

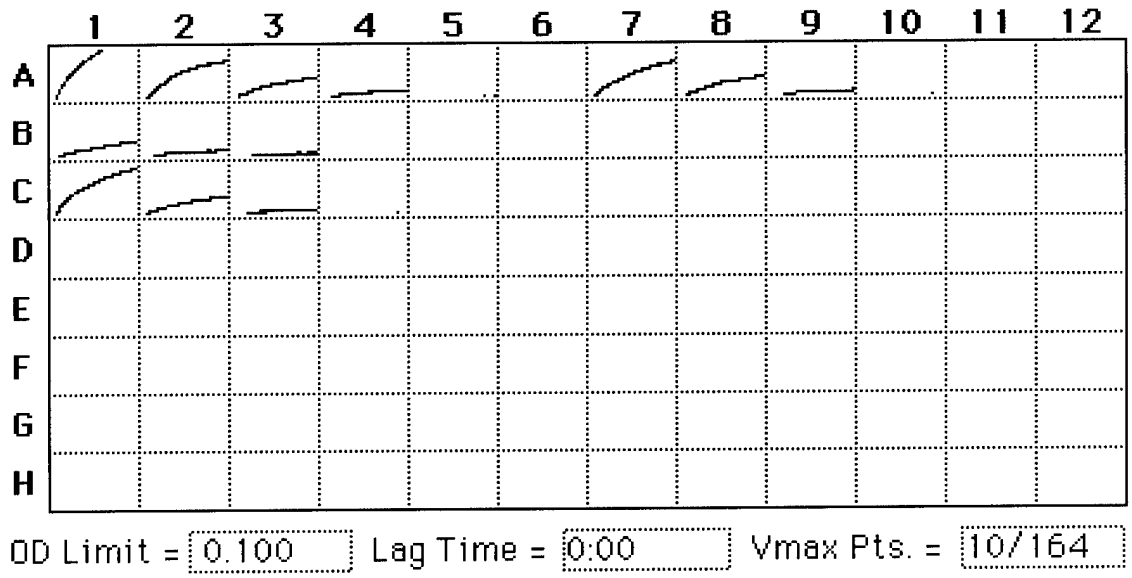


Figure 21. Trypsin Activation of Prostromelysin at 4.16 uM Trypsin. A colormetric assay for the amount of active stromelysin was used to monitor the activation of prostromelysin. Each box represents the time course of color development for one well in a microtiter plate. The substrate is Ac-Pro-Leu-Gly-[2-mercapto-4-methyl-pentanoyl]-Leu-Gly-0Et. Reactions were at room temperature for 30 minutes. The OD represented by the box height is 0.1.

Row A. Mature form of stromelysin
- Column 1: 5.0 ug/ml stromelysin
- Column 2: 2.5 ug/ml stromelysin
- Column 3: 1.2 ug/ml stromelysin
- Column 4: 0.6 ug/ml stromelysin
- Column 5: 0.3 ug/ml stromelysin

Row B. Prostromelysin untreated
- Column 1: 48 ug/ml prostromelysin
- Column 2: 24 ug/ml prostromelysin
- Column 3: 12 ug/ml prostromelysin
- Column 4: 6 ug/ml prostromelysin
- Column 5: 3 ug/ml prostromelysin

Row C. Prostromelysin treated for 30 minutes at 37C with 4.16 uM trypsin
- Column 1: 48 ug/ml prostromelysin
- Column 2: 24 ug/ml prostromelysin
- Column 3: 12 ug/ml prostromelysin
- Column 4: 6 ug/ml prostromelysin
- Column 5: 3 ug/ml prostromelysin

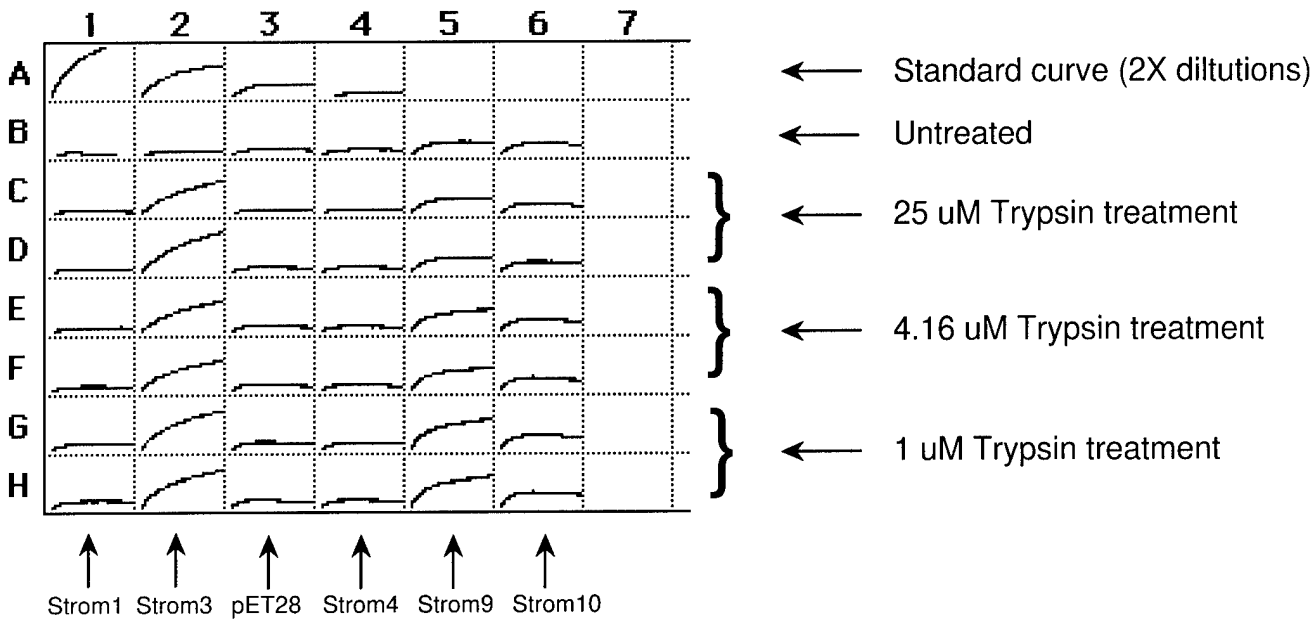


Figure 22. Trypsin Activation of Prostromelysin Made by Our Clones in Crude Lysates. A colormetric assay for the amount of active stromelysin was used to monitor the activation of prostromelysin. Each box represents the time course of color development for one well in a microtiter plate. The substrate is Ac-Pro-Leu-Gly-[2-mercapto-4-methyl-pentanoyl]-Leu-Gly-0Et. Reactions were at room temperature for 30 minutes. The OD_{405} represented by the box height is 0.1. Rows B through H all have lysate samples from the same clone.

Row A: Serial two-fold dilutions of mature stromelysin starting at 5 ug/ml
Row B: Lysates samples from six different clones UNTREATED
Rows C,D: Lysates samples from six different clones treated with 25 uM trypsin
Rows E,F: Lysates samples from six different clones treated with 4.16 uM trypsin
Rows G,H Lysates samples from six different clones treated with 1 uM trypsin

Strom3 clearly has activatable stromelysin activity (more color in C-H).
Strom9 has a low level of activatable stromelysin activity (more color in G,H).

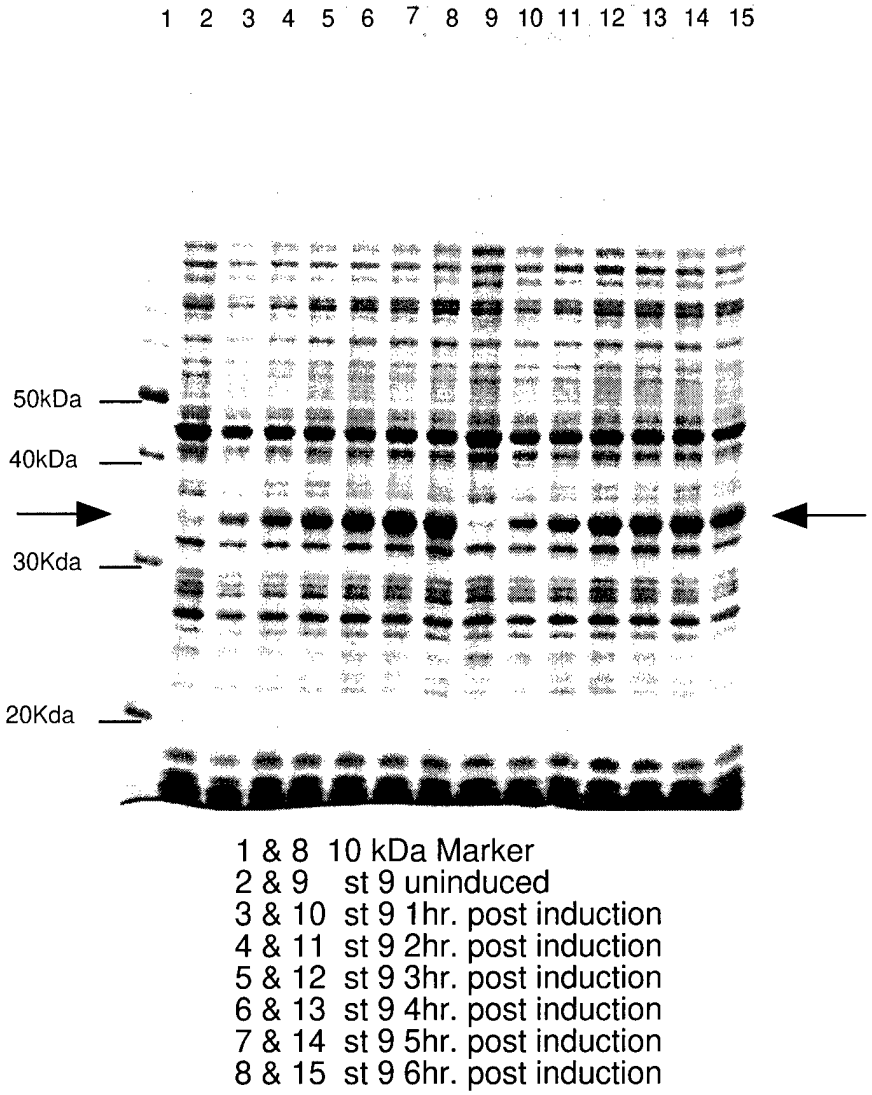
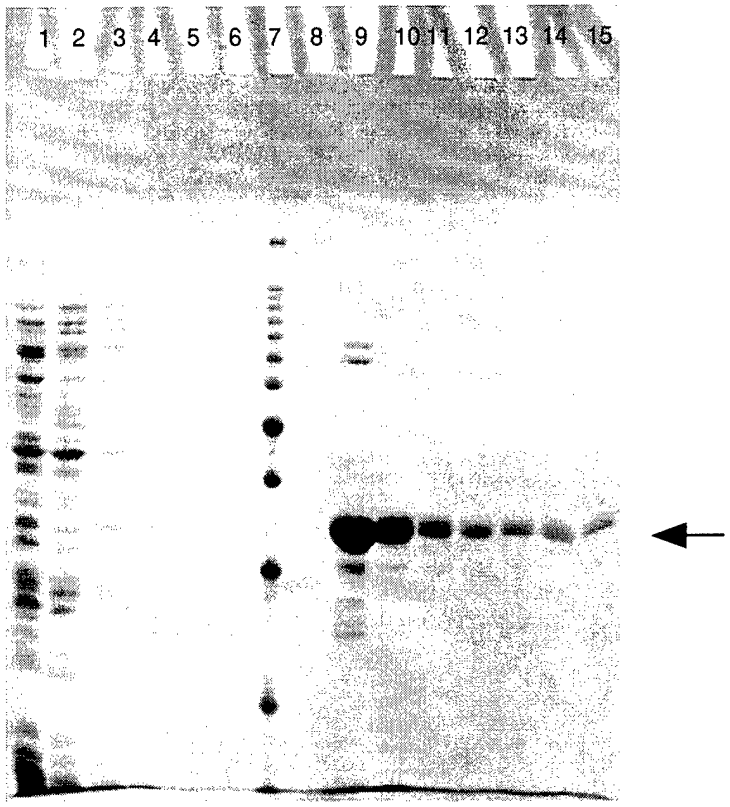


Figure 23. Induction of the Prostromelysin Strom9 Clone in *E.coli*. Duplicate cultures of the prostromelysin clone, strom9, were grown and induced for prostromelysin expression. Aliquots were taken at 1 hour intervals after induction and lysed. Cleared lysate solutions were then analyzed on a 12% polyacrylamide-1% SDS electrophoresis gel and stained with Coomassie Blue. The arrow indicates the expected size, 33kD, of the prostromelysin product.



1 non-binding flowthrough
2 Wash 1
3 Wash 2
4 Wash 3
5 Wash 4
6 Wash 5
7 10 kDa Marker
8 Elution fraction 1
9 Elution fraction 2
10 Elution fraction 3
11 Elution fraction 4
12 Elution fraction 5
13 Elution fraction 6
14 Elution fraction 7
15 Elution fraction 8

Figure 24. Analysis of Fractions from Column Purification of His-Tagged Prostromelysin. Cleared lysate from 1250 ml of culture was subjected to nickel-column fractionation to purify the his-tagged prostromelysin. Fractions were analyzed on a 12% polyacrylamide-1% SDS electrophoresis gel and stained with Coomassie Blue. The arrow indicates the expected size, 33kD, of the prostromelysin product.

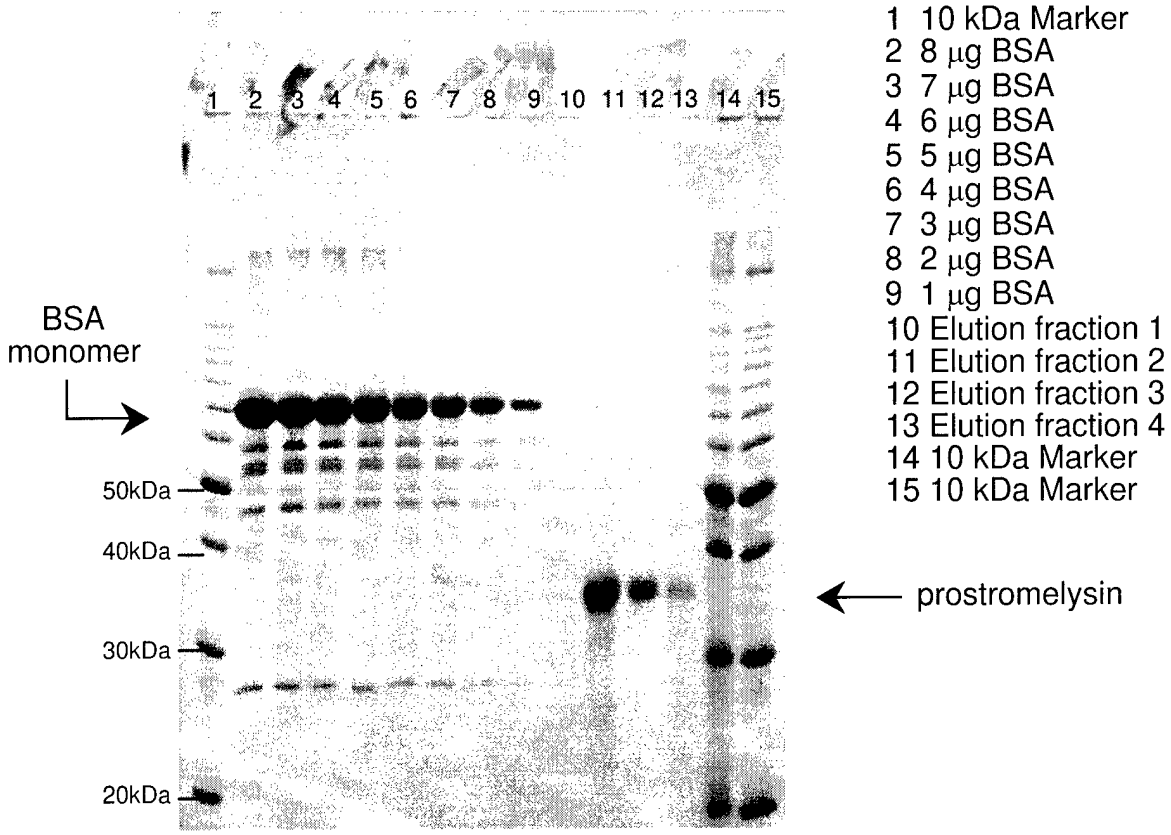


Figure 25. Quantitation of His-tagged Prostromelysin Yield from Nickel-Column. Various amounts of bovine serum albumin (BSA) were run on a 12% polyacrylamide-1% SDS electrophoresis gel to serve asstandards to quantiate the amount of prostromelysin in the elution fractions. A software package (Molecular Analyst from BioRad) was used to integrate the amount of Coomassie Blue stain in each band. The arrows indicates the locations of the prostromelysin and the BSA monomer bands.

Input Mixture	Panning Round	Input Phage Concentration	Input Ratio white:blue	Output Ratio	Enrichment Factor
Peglin #1 (W) m663 (B)	1	2.6×10^{9}	> 1:25	> 1:1	> 25
Peglin #1 (W) m663 (B)	2	2.4×10^{7}	1:7	> 1:1	> 7
Peglin #1 (W) m663 (B)	1	8.4×10^{10}	1:167	> 2:1	> 334
Peglin #1 (W) m663 (B)	2	1.9×10^{8}	> 1:200	>2.3:1	> 466

Table 1. Binding of Circularly Permuted Eglin (Peglin) to Subtilisin. The input mixture consisted of a peglin isolate that makes white plaques on indicator plates and an M13 phage wthout any binding epitope that makes blue paques on the indicator plates. The mixture was put through the panning process once, an aliquot taken and the remainder passed through the panning process again. Panning was done in microtiter plate wells coated with subtilisin. The output ratio is based onthe blue/white ratio of plaques.

Input Mixture	Papain Status	Input Ratio (Blue:White)	Ave Input Ratio (B:W)	Output Ratio (B:W)	Ave Output Ratio (B:W)	Fold Enrichment (White)
Teglin (W) m663 (B)	active	45.0:1 235.0:1	140.0:1	29.0:1	29.0:1	4.8
Teglin (W) m663(B)	inactive	18.0:1 68.0:1	42.0:1	13.0:1 11.0:1	12.0:1	3.5
Teglin-papain 13 a (W) m663 (B)	active	5.0:1 8.0:1 4.9:1	6.0:1	0.8:1 0.5:1	0.7:1	9.2
Teglin-papain 13 a (W) m663 (B)	inactive	22.0:1* 4.0:1 8.0:1	11.0:1* 6.0:1	0.3:1 0.3:1	0.3:1	37.0* 20.0
Teglin-papain 13b (W) m663 (B)	active	1.6:1 4:1:1 1.1:1	2.2:1	0.3:1 0.2:1	0.2:1	8.8
Teglin-papain 13 b (W) m663 (B)	inactive	1.4:1 9:1:1 1.3:1	2.3:1	0.2:1 0.2:1	0.2:1	14.0
m666 (W) m663 (B)	active	1.3:1 0.5:1 1.2:1	1.0:1	0.9:1 0.7:1	0.8:1	1.2
m666 (W) m663 (B)	inactive	0.9:1 1.9:1	1.4:1	0.7:1 1.8:1	1.3:1	1.1

* *Difficult to interpret due to low # of white input phage.*

Table 2. Phage Displaying Teglin Variants Binding to Papain. Phage mixtures were assayed in 96-well microtiter plates in wells coated with papain. The papain was active or inactivated by EDTA as indicated in the status column. Inactivated papain is expected to hav ethe native conformation but be unable to cleave weakly binding phage. The ratios of the various types of phages were determined by plating diltutions of the phage on x-gal indicator plates. The m666 and m663 phagesdo not have any display peptides and serve as controls for the enrichment experiment.

ISOLATE	IP Conc	IP Ratio (B:W)	OP Conc	OP Ratio (B:W)	OP_{ratio}/IP_{ratio}
L1	$3.5x10^{11}$	.9:1	$4.1x10^{6}$	1.2:1	1.3
L2	$2.7x10^{11}$	1.5:1	$1.8x10^{6}$	0 .3:1	0.2
L3	$5.3x10^{11}$	4.3:1	$1.8x10^{6}$	0.3:1	0.1
L4	$2.4x10^{11}$	5:1	$2.6x10^{6}$	1.6:1	0.3
L5	$5.2x10^{11}$	3:1	$2.6x10^{6}$	0.2:1	0.1
L6	$5.4x10^{11}$	3.5:1	$2.5x10^{6}$	0 .3:1	0.1
L7	$5.6x10^{11}$	4.6:1	$1.3x10^{6}$	0 .4:1	0 .1
L8	$3.4x10^{11}$	1.1:1	$2.8x10^{6}$	1.5:1	1.4
L9	$8.1x10^{11}$	2.1:1	$2.7x10^{6}$	0.3:1	0.1
L10	$8.9x10^{11}$	3.5:1	$3.1x10^{6}$	0.3:1	0 .1
L11	$2.5x10^{11}$	1.8:1	$2.5x10^{6}$	1.7:1	0.9
L12	$6.5x10^{12}$	4.5:1	$1.3x10^{6}$	0.3:1	0.1
L13	$5.6x10^{11}$	1.9:1	$4.2x10^{7}$	111.0:1	58.0
L14	$3.1x10^{11}$	1.8:1	$2.5x10^{7}$	81.0:1	45.0
L15	$8.0x10^{12}$	6.5:1	$1.2x10^{6}$	0.4:1	0.05
L16	$4.5x10^{11}$	4.6:1	$1.9x10^{7}$	11.0:1	2.4
L17	$9.2x10^{11}$	3:1	$2.7x10^{6}$	0 .3:1	0.1
L18	$4.9x10^{11}$	2.5:1	$1.3x10^{6}$	0 .4:1	0.1
L19	$1.4x10^{11}$	.4:1	$1.6x10^{6}$	0.4:1	1.0
L20	$7.2x10^{11}$	5.5:1	$1.5x10^{6}$	0 .2:1	0.04
L21	$1x10^{12}$	2.6:1	$2.5x10^{6}$	0.4:1	0.2
L22	$7.8x10^{11}$	2.9:1	$2.0x10^{6}$	0.5:1	0.2
L23	$4.5x10^{11}$	6.5:1	$1.2x10^{6}$	0.3:1	0.05
L24	$4.8x10^{11}$	2.7:1	$.9x10^{6}$	0.6:1	0.2
L25-L33	...	...	...	...	< 2.0
L34	$7.1 x10^{10}$	10:1	$1.5 x 10^{7}$	21.0:1	2:1

Table 3. Enrichment Factors After Panning for Isolates from the Teglin-Papain Library. Individual clones from the teglin library containing a degenerate papain binding site were picked after four rounds of panning and tested in the enrichment assay using the m663 M13 phag e as the blue plaque producer. Isolates L25 through L33 are not shown in that they all had enrichment factors below 2.

Individual Binders from Pan6 (unamplified) vs Background
Scheme:Pan/amplify x6

	Background OD_{405}	Tested	≥2x	3x	4x	5x	6x	≥10x
Alcohol dehydrogenase	.069	88	25	39	10	2	0	0
Aldolase	067	92	62	0	0	0	0	0
Alpha Amylase	.164	94	3	0	2	1	0	0
	.070	96	39	12	6	1	1	3
Catalase	.079	92	8	1	3	0	0	53
Enolase	.065	90	33	5	0	1	0	0
Hexokinase	.106	90	2	0	0	0	0	0
L-lactate dehydrogenase	.169	3	0	0	0	0	0	0
Ribonuclease-A	.070	92	15	61	15	0	0	1

Table 4. Binders to the General Enzyme Panel. Individual clones from a phage populationdisplaying randomized free peptides after six rounds of panning were tested for binding to the eight enzymes in our panel using an ELISA assay to detect phage binding. Listed are the number of clones which showed binding at the indicated levels above background.

Appendix.

Manuscripts: Published and Submitted

1. Waldner, JC, Lahr, SJ, Edgell, MH, and Pielak, GJ. Effect of a polyhistidine terminal extension on eglin c stability. *Anal. Biochem.* **263** 116-118 (1998)

2. Waldner, JC, Lahr, SJ, Edgell, MH, and Pielak, GJ. Nonideality and protein thermal denuration. Biopolymers 49 471-479 (1999)

3. Lahr, SJ, Broadwater, A, Carter, CW, Jr., Collier, ML, Hensley, L, Waldner, JC, Pielak, GJ, and Edgell, MH. Patterned library analysis: a method for the quantitative assessment of hypotheses concerning the determinants of protein structure. PNAS (submitted 1999)

Effect of a Polyhistidine Terminal Extension on Eglin c Stability

Jennifer C. Waldner,* Steven J. Lahr,†
Marshall Hall Edgell,‡ and Gary J. Pielak*,[1]

**Department of Chemistry, †Department of Biochemistry & Biophysics, and ‡Department of Microbiology and Immunology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599*

Received May 20, 1998

Polyhistidine terminal extensions (his tags) are attached to recombinant proteins to facilitate purifica-

[1] To whom correspondence should be addressed. Fax: 966-3675. E-mail: gary_pielak@unc.edu.

ANALYTICAL BIOCHEMISTRY **263,** 116–118 (1998)
ARTICLE NO. AB982808
0003-2697/98 $25.00

tion because they bind resin-immobilized divalent cations, most commonly, Ni^{2+} (1). The effects of his tags on protein stability are idiosyncratic and rarely reported (2, 3), so we set out to determine the effect of a six-residue N-terminal tag on the thermal stability of the serine protease inhibitor eglin c (4). The tag does not effect eglin c stability.

Circular dichroism-detected thermal denaturations were performed between pH 1.5 and 3.3 for both proteins. The denaturation reaction is reversible and well fit by a two-state model (5). Values for T_m (the temperature at which the thermal transition is half complete) and ΔH_m (the vant Hoff denaturation enthalpy at T_m) are given in Table 1. The change in heat capacity upon denaturation, ΔC_p, was determined by varying the pH (5). Plots of T_m versus ΔH_m yield ΔC_p values for the wild-type protein and the tagged variant of 0.84 ± 0.07 and 1.13 ± 0.13 kcal mol^{-1} K^{-1}, respectively. The uncertainties were calculated from unweighted linear least squares fitting. ΔG_D (the free energy of denaturation) at each pH was calculated using a modified form of the integrated Gibbs–Helmholtz equation (6):

$$\Delta G_{D,T} = \Delta H_m\left(1 - \frac{T}{T_m}\right) - \Delta C_p\left[(T_m - T) + T\ln\left(\frac{T}{T_m}\right)\right].$$

The uncertainty in ΔG_D was estimated by applying propagation of error analysis to the equation (5).

As shown in Fig. 1 and Table 1, the tag does not affect eglin c stability. ΔG_D values at each pH are within the experimental uncertainty for each protein. A similar result is observed for the eglin c homologue, chymotrypsin inhibitor II (CI2), which possesses a natural 19-residue N-terminal tail (7, 8). Deletion of this tail has a negligible effect on stability (9). Presumably, the tails do not affect stability because they are unstructured in both the native and denatured states.

In summary, the his tag does not affect the thermal stability of eglin c. This finding will facilitate our studies of eglin c variants obtained using a high-throughput activity screen.

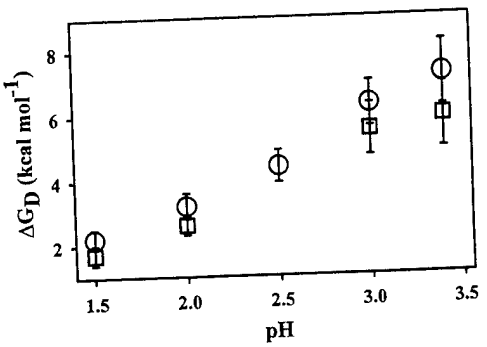


FIG. 1. ΔG_D-versus-pH plot for wild-type (○) and his-tagged (□) eglin c at 25.0°C. The error bars were obtained as described in the text.

TABLE 1

Thermodynamic Parameters for Denaturation of Eglin c and His-Tagged Eglin c

	$T_m \pm 1.6^a$ (°C)		$\Delta H_m \pm 6.3^a$ (kcal/mol)	
pH	Wild-type	His tag	Wild-type	His tag
1.5[b]	45.1	45.9	45.3	40.8
2.0[b]	47.3	48.4	43.8	45.8
2.5[c]	54.6	56.4	51.2	56.2
3.0[e]	62.2	63.4	57.4	63.0
3.3[c]	69.0	69.0	64.3	69.2

[a] Uncertainties are the standard deviations of the mean from four repetitions at pH 3.0 and are representative of the uncertainties at other pH values.

The average values from [b]two, [c]three, [d]four, or [e]five trials.

Materials and Methods

Expression and purification. The eglin c gene was inserted into the pET17b vector (Novagen) and expressed using the *Escherichia coli* strain BL21(DE3)pLysS (10). After lysis, cell extracts were brought to pH 3.0 and insoluble proteins removed by centrifugation. Eglin c was purified from the supernatant by using a Sephadex G-75 gel-filtration column equilibrated in 50 mM glycine–HCl, pH 3.0.

Construction and purification of tagged eglin c. The eglin c gene was removed from pET17b and inserted into the N-terminal His tag-containing vector pET28a (Novagen). Transformants were induced with 1 mM isopropyl β-D-thiogalactopyranoside. Purification was completed using Novagen's pET His-Tag system.

CD-detected thermal denaturation. Data were acquired with an Aviv Model 62DS spectropolarimeter equipped with a five-position sample chamber. Experiments were performed in 50 mM glycine–HCl buffer using protein concentrations of 60 to 70 μM. The ellipticity at 227 nm was followed from 5 to 90°C at 1°C intervals, with ≈6.0 min between temperatures. Reversibility was checked by returning the samples to the initial temperature and repeating the experiment (5).

Acknowledgments. This work was supported by the U.S. Army Medical Research and Material Command under DAMD17-94-J-4270 and by NIH Grant GM42501. J.C.W. was partially supported by a GAANN fellowship from the U.S. Department of Education. We thank A. Broadwater and L. Hensley for technical assistance and the Pielak and Edgell groups for helpful discussions.

REFERENCES

1. Hengen, P. N. (1998) *Trends Biochem. Sci.* **20,** 286.
2. Deutscher, S. L., Crider, M. E., Ringbauer, J. A., Komissarov, A. A., and Quinn, T. P. (1996) *Arch. Biochem. Biophys.* **333,** 207–213.

3. Ledent, P., Duez, C., Vanhove, M., Lejeune, A., Fonze, E., Charlier, P., Rhazi-Filali, F., Thamm, I., Guillaume, G., Samyn, B., Devreese, B., Van Beeumen, J., Lamotte-Brasseur, J., and Frere, J. M. (1997) *FEBS Lett.* **413,** 194–196.
4. Seemüller, U., Meier, M., Ohlsson, K., Müller, H. P., and Fritz, H. (1977) *Hoppe Z. Physiol. Chem.* **358,** 1105–1107.
5. Cohen, D. S., and Pielak, G. J. (1994) *Protein Sci.* **3,** 1253–1260.
6. Elwell, M. L., and Schellman, J. A. (1977) *Biochim. Biophys. Acta* **494,** 367–383.
7. Kjaer, M., Ludvigsen, S. J., Sorenson, O. W., Denys, L. A., Kindtler, J., and Poulsen, F. M. (1987) *Carlsberg Res. Commun.* **52,** 327–354.
8. Kjaer, M., and Poulsen, F. M. (1987) *Carlsberg Res. Commun.* **52,** 355–362.
9. Jackson, S. E., Moracci, M., el Masry, N., Johnson, C. M., and Fersht, A. (1993) *Biochemistry* **32,** 11259–11269.
10. Grodberg, J., and Dunn, J. J. (1988) *J. Bacteriol.* **170,** 1245–1253.

Nonideality and Protein Thermal Denaturation

Jennifer C. Waldner[1]
Steven J. Lahr[2]
Marshall Hall Edgell[3]
Gary J. Pielak[1]

[1] *Department of Chemistry, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3290*

[2] *Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7260*

[3] *Department of Microbiology and Immunology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7290*

Received 27 May 1998; accepted 27 October 1998

Abstract: *We studied the thermal denaturation of eglin c by using CD spectropolarimetry and differential scanning calorimetry (DSC). At low protein concentrations, denaturation is consistent with the classical two-state model. At concentrations greater than several hundred μM, however, the calorimetric enthalpy and the midpoint transition temperature increase with increasing protein concentration. These observations suggested the presence of intermediates and/or native state aggregation. However, the transitions are symmetric, suggesting that intermediates are absent, the DSC data do not fit models that include aggregation, and analytical ultracentrifugation (AUC) data show that native eglin c is monomeric. Instead, the AUC data show that eglin c solutions are nonideal. Analysis of the AUC data gives a second virial coefficient that is close to values calculated from theory and the DSC data are consistent with the behavior expected for nonideal solutions. We conclude that the concentration dependence is caused by differential nonideality of the native and denatured states. The nondeality arises from the high charge of the protein at acid pH and is exacerbated by low buffer concentrations. Our conclusion may explain differences between van't Hoff and calorimetric denaturation enthalpies observed for other proteins whose behavior is otherwise consistent with the classical two-state model.*
 Biopoly 49: 471–479, 1999

Correspondence to: Gary J. Pielak; email: gary_pielak@unc.edu
Contract grant sponsor: U.S. Army Medical Research and Material Command (U.S. Army) and NIH
Contract grant number: DAMD17-94-J-4270 (U.S. Army) and GM42501 (NIH)
Biopolymers, Vol. 49, 471–479 (1999)

CCC 0006-3525/99/060471-09

Keywords: analytical ultracentrifugation; CD; spectropolarimetry; differential scanning calorimetry; Donnan equilibria; eglin c; nonideality; protease inhibitor; protein stability; protein thermodynamics

INTRODUCTION

Considerable information about protein stability has been obtained by monitoring thermal denaturation. However, these studies generally assume that protein solutions are ideal. Because stability data are often obtained at high protein concentrations, this assumption may not always be valid. To investigate this possibility, we studied eglin c thermal denaturation with CD spectropolarimetry and differential scanning calorimetry (DSC) as a function of protein concentration. We also used analytical ultracentrifugation (AUC) to investigate the possibility of eglin c aggregation or thermodynamic nonideality.

Monitoring the disappearance of CD-detected secondary structure with increasing temperature is perhaps the most common method for measuring protein stability. One drawback of this method is that the denaturation mechanism must be assumed to obtain the thermodynamic parameters. Usually, a two-state model, where only the native and denatured states are significantly populated, is assumed and the parameters are obtained indirectly via van't Hoff analysis.[1]

On the other hand, DSC provides a direct measure of the denaturation enthalpy ΔH_D.[2] DSC data can also be analyzed via van't Hoff analysis. Information can then be gained by comparing ΔH_D from van't Hoff analysis (ΔH_{vH}) to ΔH_D measured calorimetrically (ΔH_{cal}). If ΔH_{vH} equals ΔH_{cal}, unfolding is consistent with the proposed model. The observation that ΔH_{vH} is less than ΔH_{cal} is usually interpreted as evidence for intermediates. We show that other interpretations are valid.

Information is also gained about the unfolding reaction by studying the protein concentration dependence of the thermodynamic parameters. Typically, native state aggregation is inferred if T_m, the temperature at which the thermal transition is half over, increases with increasing protein concentration.[3] However, another source of protein concentration dependence, thermodynamic nonideality, is rarely investigated. Both aggregation and nonideality can be quantified with AUC.

We studied the thermal denaturation of *Hirudo medicinalis* eglin c,[4] a small [calculated molecular weight (MW_{calc}),8.2 kD], single domain protein that lacks metal-ion binding sites and disulfides. Eglin c has several properties that make it an attractive candidate for biophysical analysis. It can be expressed in *Escherichia coli* in hundred-milligram quantities, and its three-dimensional structure is known to high resolution.[5–10] Additionally, there is a wealth of information about its denaturation. Bae and Sturtevant[11] studied the denaturation of eglin c between pH 1.5 and 11.0 with DSC. Also, the eglin c homologue, chymotrypsin inhibitor II (CI2) [1.68 Å root mean square (rms) deviation[6]] denatures by a two-state mechanism.[12–14]

Here we report the thermodynamic parameters for eglin c denaturation from CD and DSC data. The results provoked us to examine eglin c aggregation and nonideality with AUC. We report a detailed analysis of eglin c thermal denaturation, including a model that explains the concentration dependence of the thermodynamic parameters.

MATERIALS AND METHODS

Protein Purification and Expression

The eglin c gene was synthesized and inserted into the pET17b plasmid (Novagen), utilizing the *Eco*RI and *Nde*I restriction enzyme sites, to produce the construct ME007. The amino acid sequence used here is identical to that described by Rink et al.[15] For protein expression, ME007 was transformed into the *E. coli* strain BL21(DE3)pLysS ($F^- ompT\ r_B^- m_B^-$).[16] A single colony was used to inoculate 50 mL of Luria broth (LB) containing 100 μg mL^{-1} carbenicillin plus 34 μg mL^{-1} chloramphenicol and grown to an $OD_{600\ nm}$ of 0.6 with shaking at 37°C. The cells were then pelleted at 5000 x g for 15 min and used to inoculate 1 L of fresh LB and antibiotics. The culture was incubated at 37°C, with shaking, to an $OD_{600\ nm}$ of 0.8. The cells were pelleted as described above, resuspended in 1 L of LB containing fresh antibiotics, and induced with 0.4 m*M* isopropyl β-D-thiogalactopyranoside. Cells were harvested by centrifugation 3 h after induction. The pellet was stored at −70°C overnight.

Cells were lysed by using the freeze–thaw method[17,18] and suspended in 100 mL 50 m*M* Tris HCl, pH 8.5, 5 m*M* EDTA. The lysate was treated with 10 μg μL^{-1} deoxyribonuclease for 1 h at 37°C followed by centrifugation at 10,000× *g* for 15 min. The supernatant was removed and dialyzed overnight against 50 m*M* glycine-HCl, pH 3.0, followed by centrifugation at 10,000 x g for 15 min. Eglin c was purified from the supernatant by using a Sephadex G-75 column equilibrated in 50 m*M* glycine-HCl, pH 3.0. Fractions analyzed by sodium dodecyl sulfate–polyacrylamide gel electrophoresis showed eglin c as a single band when stained with Coomassie brilliant blue. A typical yield is 50 mg of pure eglin c per liter of culture.

Calorimetry and Determination of Eglin c Concentration

Experiments were performed in 50 m*M* glycine-HCl buffer at pH 1.5, 2.0, 2.5, 3.0, and 3.3. Under these conditions, the protein is soluble, and no aggregation is observed upon heating. Samples were prepared by overnight dialysis and degassed immediately prior to loading the sample cell. We used a glycine buffer because the ionization enthalpy of its carboxyl groups is small ($\sim$1 kcal mol^{-1})[19] and, therefore, the pH does not change significantly over the temperature range used. Calorimetry was performed with a MCS-Differential Scanning Calorimeter from MicroCal, Inc. (Northampton, MA). The dialysis buffer was placed in both the sample and reference compartments, and several thermograms were acquired to establish a baseline. The sample buffer was then replaced with buffer containing eglin c. Samples were scanned from 10 to 90°C at a scan rate of 60°C h^{-1}, unless otherwise noted.

The molar extinction coefficient of eglin c was determined by quantitative amino acid analysis and the difference spectrum method.[20,21] The values agree, yielding an ϵ_{280} of 1.24×10^4 M^{-1}cm^{-1}. The protein concentration was determined by measuring the absorbance in a 0.1 cm path-length cuvette so that sample dilution was unnecessary. Except for concentration-dependence experiments, protein concentrations of 100–130 μM were used.

Analysis of Calorimetric Data

The appropriate buffer baseline was subtracted, and the data were converted to molar excess heat capacity by using the protein concentration and the cell volume (1.21 mL). Thermograms were analyzed with MicroCal Origin software. To obtain T_m, ΔH_{vH},and ΔH_{cal}, a baseline through the pre- and posttransitions was subtracted to eliminate changes in heat capacity between the native and denatured states. Uncertainties were determined by repetition of experiment at pH 3.0. Data also were analyzed with a model that includes reversible dimerization of the native and denatured states.[22]

CD Spectropolarimetry

Data were acquired with an Aviv model 62DS CD spectropolarimeter equipped with a five-position sample chamber and a Peltier effect thermoelectric temperature controller. Experiments were performed in 50 m*M* glycine-HCl at 1°C intervals with 6.0 min between temperatures. The protein concentration was 60–70 μM. The ellipticity at 227 nm was followed from 5 to 70°C at pH 1.5 and 2.0 and from 30 to 90°C at pH 2.5, 3.0, and 3.3. Data were fit to a two-state model with linear baselines.[1] Uncertainties were determined by repetition of experiment. Reversibility was checked by returning the heated samples to the initial temperature and repeating the denaturation experiment.

Determining ΔC_p

In principle, ΔC_p can be obtained from one DSC thermogram. In practice, however, this method is unreliable.[22–25] Because carboxylate enthalpies of ionization are small, ΔC_p can be determined more precisely by measuring the temperature dependence of ΔH_D at several pH values. Because the calorimetry data were analyzed directly and via van't Hoff analysis, both ΔH_{cal} and ΔH_{vH} were used to obtain ΔC_p.

Determining $\Delta\nu$, $\Delta H_{D,T}$, ΔS_D, and ΔG_D

The number of protons bound to the denatured state minus the number bound to the native state, $\Delta\nu$, was estimated with[26,27]

$$\Delta\nu = \frac{\Delta H_{vH}}{2.3RT_m^2}\frac{\delta T_m}{\delta pH} \tag{1}$$

$\delta T_m/\delta$pH was obtained by fitting a second-order polynomial to a T_m vs pH plot and evaluating the first derivative at each T_m. $\Delta H_{D,T}$ at each pH was calculated using

$$\Delta H_{D,T} = \Delta H_{vH} + \Delta C_p(T - T_m) \tag{2}$$

ΔS_D was calculated using

$$\Delta S_D = \Delta S_{vH} + \Delta C_p \ln\left(\frac{T}{T_m}\right) \tag{3}$$

where ΔS_{vH} is $\Delta H_{vH}/T_m$. ΔG_D was calculated by combining Eqs.(2) and (3)[28]:

$$\Delta G_{D,T} = \Delta H_D\left(1 - \frac{T}{T_m}\right) - \Delta C_p\left[(T_m - T) + T\ln\left(\frac{T}{T_m}\right)\right] \tag{4}$$

Analytical Ultracentrifugation

Sedimentation equilibrium experiments were performed at 300 K with a Beckman XLA analytical ultracentrifuge in the Macromolecular Interactions Facility at University of North Carolina, Chapel Hill. Protein concentrations ranged from 13 to 54 μM. Experiments were conducted in 50 m*M* glycine-HCl buffer, pH 3.0. The samples were centrifuged at 22,000, 27,000, 40,000, and 45,000 rpm. The absorbance at 280 nm as a function of radial distance was measured every hour. The system was assumed to be at equilibrium when successive scans overlaid. Assuming that any nonideality can be described by the second virial coefficient, the equilibrium protein concentration distribution at radius r, $c(r)$, is given by

$$c(r) = c_0\exp(\sigma\xi - 2B[c(r)]) \tag{5}$$

where σ is the reduced monomer molecular weight,

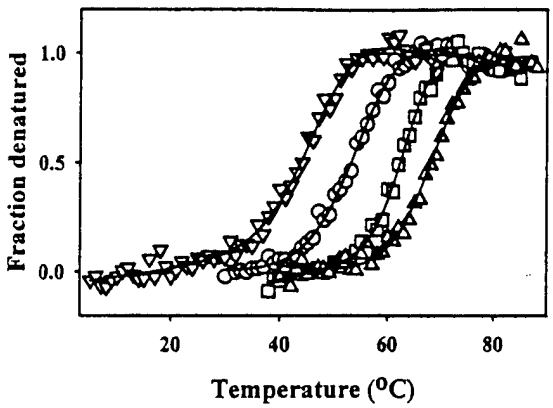


FIGURE 1 Fraction denatured vs temperature plot and the fits from nonlinear least-squares fitting for CD-detected thermal denaturation at pH (▽)1.5, (○) 2.0, (□) 3.0, and (△) 3.3. Every third point is shown.

$$\sigma = \frac{MW(1 - \nu\rho)\omega^2}{RT} \qquad (6)$$

c_0 is the initial protein concentration, ν is the partial specific volume (from the weighted average of the individual amino acids[19]), ρ is the buffer density, ω is the angular velocity, $\xi = (r^2 - r_o^2)/2$, r_0 is the reference radial position, and B is the second virial coefficient. Analysis of the data with B set to zero gives the apparent molecular weight (MW_{app}). See Laue[29] for a more detailed description of these parameters. Data were fit with the computer program NONLIN.[29,30] Because $c(r)$ is measured in absorbance units (A), $B \cdot MW_{calc}$ from NONLIN has units of A^{-1}. Beers law along with a path length of 1.20 cm was used to convert A^{-1} to M^{-1}, the standard units for $B \cdot MW_{calc}$.

RESULTS

CD-Detected Thermal Denaturation

Representative denaturation curves at four pH values are shown in Figure 1 along with fits to a two-state model. Greater than 95% of the signal returns upon cooling and parameters from successive scans are within experimental uncertainty. These observations show that denaturation is reversible. T_m and ΔH_{vH} at the five pH values are given in Table I. From the temperature dependence of ΔH_{vH}, ΔC_p is 0.84 ± 0.07 kcal mol^{-1} K^{-1} (Figure 2).

As shown in Figure 3, $\Delta\nu$ is −2.3 at pH 3.0. Using the data in Table I [Eq. (4)], and the above-mentioned ΔC_p, $\Delta G_{D,300K}$ is 4.2 kcal mol^{-1} at pH 3.0. Both $\Delta\nu$ and $\Delta G_{D,300K}$ decrease with decreasing pH.

Calorimetry

The thermograms are symmetric at all protein concentrations[31] and show a positive value for the change in heat capacity upon denaturation ΔC_p. For any specific set of conditions, all the scans are nearly superimposable and thermodynamic parameters from successive scans are within experimental uncertainty, indicating that denaturation is reversible.

T_m, ΔH_{cal}, and ΔH_{vH} were obtained using MicroCal software. A typical data set and fit are shown in Figure 4. The parameters obtained at five pH values are given in Tables II and III for experiments performed at low protein concentration. The scan-rate independence of T_m and $\Delta H_{vH}/\Delta H_{cal}$ shows that we

Table I Thermodynamic Parameters From CD Data[a]

pH	T_m ± 1.6[b] (°C)	ΔH_{vH} ± 6.3[b] (kcal mol^{-1})
1.5[c]	45.1	45.3
2.0[c]	47.3	43.8
2.5[d]	54.6	51.2
3.0[e]	62.2	57.4
3.3[d]	69.0	64.3

[a] ΔH_{vH} is reported at T_m. Experiments were performed in 50 mM glycine-HCl, pH 3.0, using 60–70 μM eglin c.

[b] Uncertainties are the standard deviation of the mean from five repetitions at pH 3.0. Uncertainties at other pH values are assumed to be those of pH 3.0.

[c] The average values from two repetitions.

[d] The average values from three repetitions.

[e] The average values from five repetitions.

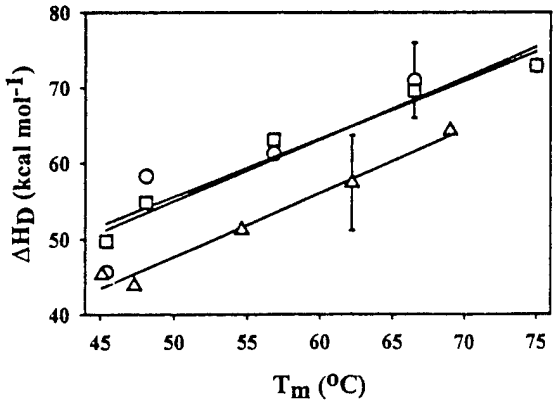


FIGURE 2 Plots of (○) ΔH_{cal}, (□) ΔH_{vH}(DSC), and (△) ΔH_{vH}(CD) vs T_m. The average value for each pH is shown (Tables I and II). The error bars represent the standard deviation of the mean from repetition of experiment. Linear least-squares fitting of the CD data gives a slope of 0.84 ± 0.07 kcal mol^{-1} K^{-1} and a correlation coefficient (r^2) of 0.98. Linear least-squares fitting of the ΔH_{vH} and ΔH_{cal} data from DSC gives slopes of 0.84 ± 0.07 kcal mol^{-1} K^{-1} and 0.82 ± 0.19 kcal mol^{-1} K^{-1} and r^2 values of 0.96 and 0.86, receptively. P_r for all lines is <3.7%.

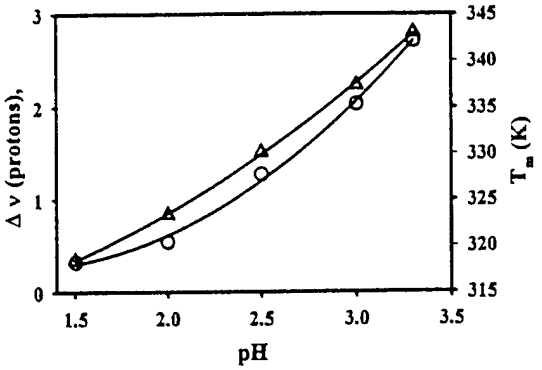


FIGURE 3 Plots of T_m (○) and $\Delta\nu$ (△) vs pH. $\Delta\nu$ was calculated by using Eq. (1) and data in Table I. The curve through the $\Delta\nu$ values has no theoretical significance.

are making equilibrium measurements. From the temperature dependence of ΔH_{vH} and ΔH_{cal} (Figure 2), ΔC_p is 0.84 ± 0.07 kcal mol^{-1} K^{-1} and 0.82 ± 0.19 kcal mol^{-1} K^{-1}, respectively. These values agree with our CD-detected values and with Bae and Sturtevant's calorimetric value.[11] Interestingly, we obtain conflicting results regarding the temperature dependence of ΔC_p. The pre- and posttransition baselines slopes are different, suggesting that ΔC_p is temperature dependent. However, ΔH_D vs T_m plots fit a straight line with a high correlation coefficient, suggesting that ΔC_p is temperature independent (Figure 2). This discrepancy is observed by others.[22,23,32]

At low protein concentrations, ΔH_{vH} and ΔH_{cal} are equal at any given pH, suggesting that denaturation is a two-state process. However, at protein concentrations >130 μM, the thermodynamic parameters become concentration dependent (Figure 5). Increasing the concentration from 80 to 600 μM increases T_m by 10°C and ΔH_{cal} by 30 kcal mol^{-1}, while μH_{vH} decreases slightly. Because ΔH_{vH} decreases and ΔH_{cal} increases, $\Delta H_{vH}/\Delta H_{cal}$ decreases with increasing protein concentration. Concentration dependence was not observed by Bae and Sturtevant from 150 to 500 μM.[11] We cannot explain this discrepancy.

Analytical Ultracentrifugation

To probe the source of the concentration dependence, sedimentation equilibrium experiments were performed as a function of protein concentration under the same buffer conditions used for stability measurements (50 mM glycine-HCl, pH 3.0). First, the data were first fit separately at each concentration. For this analysis, B in Eq. (5) was set to 0 M^{-1} and MW was allowed to vary to give a MW_{app} at each protein concentration. A plot of $MW_{app}/2$ vs (protein concentration)$^{-1}$ has a slope ($B \cdot MW_{calc}$) of $1.2 \times 10^3\ M^{-1}$. To obtain a more precise value we exploited the global fitting feature of NONLIN. Fitting the data to Eq. (5) with MW set equal to MW_{calc} (8.2 kD) gives a $B \cdot MW_{calc}$ of $1.2 \pm 0.2 \times 10^3\ M^{-1}$. A representative data set and fit are shown in Figure 6A. Panel B shows that the residuals are random and centered around zero. Ultracentrifugation experiments were also conducted at high ionic strength (50 mM glycine HCl and 0.5M NaCl). Unfortunately, these data are uninterpretable because eglin c aggregates at all protein concentrations under these conditions.

DISCUSSION

Evidence for Two-State Behavior and Against Native-State and Denatured-State Aggregation

Several observations support our conclusion that eglin c denatures by a classical two-state process at low protein concentration. First, the mean ratio of ΔH_{vH} to ΔH_{cal} from Table II is 1.01 ± 0.07. Similar results were obtained for the eglin c homologue, CI2 ($\Delta H_{vH}/\Delta H_{cal}$ = 0.98 ± 0.03.[12] Second, ΔH_{vH} from CD and DSC are within experimental uncertainty (Tables I and II). Third, the thermograms are symmetric at all protein concentrations.[31]

On the other hand, at high protein concentrations, $\Delta H_{vH} < \Delta H_{cal}$ and both ΔH_{cal} and T_m increase with increasing protein concentration. These observations are inconsistent with the classical two-state model. One explanation for an increase in T_m with increasing protein concentration is native-state association.[3] Several lines of evidence are inconsistent with aggrega-

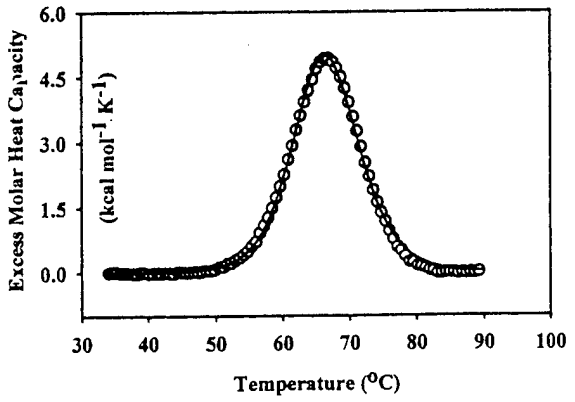


FIGURE 4 Typical thermogram obtained with 130 μM eglin c in 50 mM glycine-HCl, pH 3.0. Every other data point is shown. Baselines through the pre- and posttransition were subtracted to normalize the change in heat capacity between the end states. The curve is from nonlinear least-squares fitting.

Table II Thermodynamic Parameters From DSC Data[a]

pH	$T_m \pm 0.1$[b] (°C)	$\Delta H_{vH} \pm 0.7$[b] (kcal mol^{-1})	$\Delta H_{cal} \pm 5.0$[b] (kcal mol^{-1})	$\Delta H_{vH}/\Delta H_{cal} \pm 0.07$[c]
1.5[d]	45.4	49.7	45.6	1.09
2.0[d]	48.1	54.8	58.3	0.94
2.5[d]	56.8	63.1	61.3	1.03
3.0[e]	66.5	69.6	71.0	0.98
3.3[d]	75.0	72.9	72.9	1.00

[a] Values of ΔH_{vH} are reported at T_m. Experiments were performed in 50 m*M* glycine-HCl, pH 3.0, using 100–130 μM eglin c.

[b] Uncertainties are the standard deviation of the mean from three trials at pH 3.0. Uncertainties at other pH values are assumed to be those of pH 3.0.

[c] Uncertainties in $\Delta H_{vH}/\Delta H_{cal}$ were determined as described in the text.

[d] Average values from two repetitions.

[e] Average values from three repetitions.

tion. First, the calorimetric data are not well fit by a model that permits native- and/or denatured-state dimerization[22]:

$$N_2 \leftrightarrows 2N \leftrightarrows 2D \leftrightarrows D_2 \qquad (7)$$

The rms deviation between the data and the fit was typically 38% of the heat capacity maximum, compared to 1.5% for a two-state model. Second, fitting the AUC data to Eq. (5), with B set to 0, shows that eglin c is monomeric. As discussed below, these apparent discrepancies are a consequence of the nonideality of highly concentrated eglin c solutions at acid pH and low buffer concentration.

Eglin c Solutions are Nonideal

When the AUC data are fit individually at each concentration (with **B** = 0), MW_{app} becomes less than MW_{calc} and decreases with increasing protein concentration. This observation shows that the eglin c solutions are nonideal.[29] Nonideality arises from unfavorable interactions between the solute (protein) and the buffer, altering the protein concentration distribution.[33] Under these conditions, data should be fit to Eq. (5) with a nonzero B. Fitting the data in this way gives a $B \cdot MW_{calc}$ of $1.2 \pm 0.2 \times 10^3 M^{-1}$.

Nonideality arises from charge–charge repulsion (Donnan effects) and from excluded-volume effects.[34] The latter describes the volume available to a protein in solution and is discussed below.[35,36] The other source of nonideality is charge–charge repulsion. If the buffer ionic strength is too low, the protein con-

Table III Scan Rate Dependence of T_m and $\Delta H_{vH}/\Delta H_{cal}$[a]

Scan Rate (°C h^{-1})	T_m (±0.1°C)[b]	$\Delta H_{vH}/\Delta H_{cal}$ (±0.06)[b]
20	62.7	0.97
60	63.1	1.17
90	62.6	0.98

[a] Experiments were performed in 50 m*M* glycine-HCl, pH 3.0, 130 μM eglin c.

[b] Uncertainties were determined as described in the text.

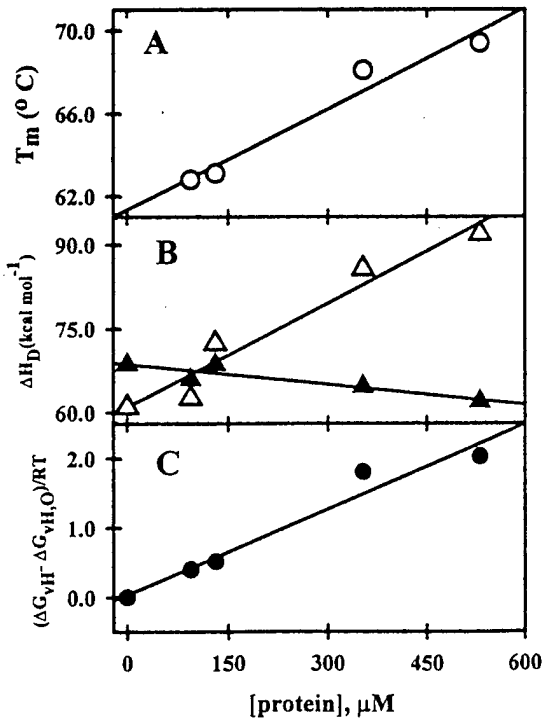


FIGURE 5 The concentration dependence of (A) T_m (○), (B) ΔH_{vH} (▲)and ΔH_{cal} (△) and (C) $(\Delta G_{vH} - \Delta G_{vH,0})/RT$ (●) at 344 K. Data were acquired at pH 3.0. Except for panel C, the curves have no theoretical significance.

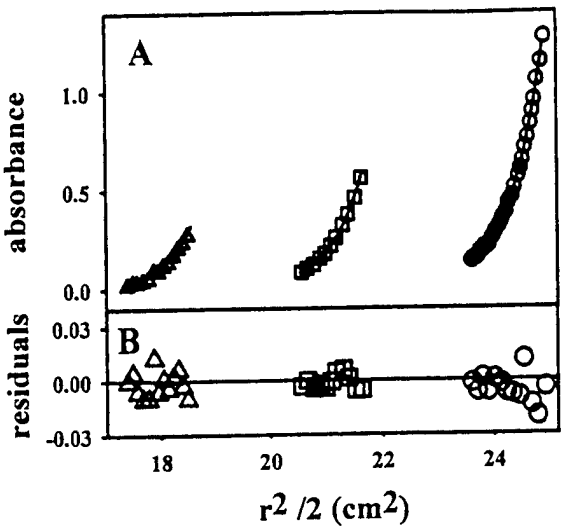


FIGURE 6 (A) Absorbance vs radial position plots from sedimentation equilibrium experiments as a function of eglin c concentration [13 (△), 27 (□), and 54 (○) μM] at 45,000 rpm and 300 K. (B) Residuals vs radial position plots. Every fourth point is shown.

centration gradient becomes linked to the buffer concentration gradient and the MW_{app} [i.e., the MW obtained from Eq. (5) when B equals zero] is less than MW_{calc} (the MW calculated from the sequence).[33] All experiments were performed at pH ≤ 3.3 in 50 mM buffer. Eglin c carries a +10 charge at pH 3 as estimated from its amino acid composition and assuming unperturbed pK values. Furthermore, the protein charge is not expected to be strongly temperature dependent because charge changes at acid pH are dominated by protein carboxylates, which, as stated above, have small ionization enthalpies.

The data support the idea that the low buffer concentration and the high protein charge bring about the observed nonideality. To test this idea, we estimated $B \cdot MW_{calc}$ for the native state by using[37,38]

$$B \cdot MW_{calc} = \frac{16\pi N R_A^3}{3} + \frac{Z^2(1 + 2\kappa R_A)}{4I(1 + \kappa R_A)^2} \qquad (8)$$

The first term describes the effect of excluded volume and the second term describes charge–charge repulsions. In Eq. (8), N is Avogadro's number, R_A is the radius (in dm for the first term and cm for the second term), Z is the protein charge (+10 for the pH 3 data), I is the molar ionic strength (0.09M at pH 3), and 6 is the inverse screening length [$=3.27 \times 10^7 \times I^{0.5}$, in cm^{-1}]. We estimated R_A by assuming eglin c is a sphere of volume $M_{app}(\nu + \delta\rho)$, where ν, the partial specific volume, δ, the solvation parameter,[34] and ρ, the solvent density, are 0.73 mL/g, 0.4 g/g, and 1.00 g/mL, respectively. The value of R_A, 1.5×10^{-7} cm, is consistent with the dimensions of the protein as estimated from examination of the crystal structures.[5–10] Using these parameters in Eq. (8), we obtain a value of $6.1 \times 10^2 M^{-1}$, half the value obtained from AUC. Using the range of values for δ compiled by Tanford[34] for globular proteins, 0.35–1.07 g/g, gives a range for the viral coefficient of $6.1 \times 10^2 M^{-1}$ to $6.5 \times 10^2 M^{-1}$. Importantly, the charge–charge term accounts for >80% of the estimated virial coefficient. In summary, the results from both experiment and theory show that the nonideality is mainly the result of charge–charge repulsion between the protein and the buffer.

Equation (8) also can be used to estimate $B \cdot MW_{calc}$ for the denatured state. Eglin c takes up ~2.3 protons upon denaturation (Figure 3), increasing Z to 12.3. To estimate a minimum value for $B_D \cdot MW_{calc}$ we assume that the denaturation-induced change in R_A has a minor effect compared to charge–charge interactions. Using this assumption, we obtain a value for $8.9 \times 10^2 M^{-1}$, a 1.4-fold increase compared to the native state. Increasing the radius, as is expected for denaturation, will further increase $\mathbf{B_D} \cdot MW_{calc}$. In summary, the nonideality of the denatured state is expected to be greater than that of the native state.

Nonideality and Calorimetric Data

We can analyze the effect of nonideality on DSC-derived parameters because second virial coefficients from AUC are equivalent to those from colligative properties and equilibrium constants.[34,39] We assume that the reaction is two state, that the shape of the thermogram reflects the progress of the reaction, and that the virial expansion can be truncated at the second coefficient. With these assumptions, the relationship between ΔG_D from van't Hoff analysis of the calorimetric data (ΔG_{vH}) and the second virial coefficients for the native (B_N) and denatured (B_D) states is

$$\frac{\Delta G_{vH} - \Delta G_{vH,0}}{RT} \approx 2([D]B_D MW - [N]B_N MW) \qquad (9)$$

where $\Delta G_{vH,0}$ is ΔG_{vH} at zero protein concentration.[40] $\Delta G_{vH,0}$ is obtained by using Eq.(4) with T_m and ΔH_{vH} extrapolated to 0M protein (Figure 5) and a ΔC_p of 0.82 kcal mol^{-1} K^{-1}. ΔG_{vH} is obtained similarly except that ΔH_{vH} and T_m are from experiments at finite protein concentrations (Figure 5). We evaluated Eq. (9) at 334 K, the T_m at 0M protein.

Near T_m, [D] ~ [N] and the slope of the plot in Figure 5 is $MW \cdot (B_D - B_N)$. The slope of the plot in Figure 5C is positive, consistent with the idea that solutions of the denatured protein are more nonideal than solutions of the native protein. Using slope from Figure 5C, the $B_N \cdot MW$ from AUC and assuming that B is temperature independent, we obtain a $B_D \cdot MW$ of $2.9 \times 10^3 M^{-1}$ using Eq. (9). This amounts to a 2.4-fold increase in $B \cdot$ MW upon denaturation ($B_D \cdot MW/B_N \cdot MW = 2.9 \times 10^3 M^{-1}/1.2 \times 10^3 M^{-1}$) compared to the 1.4-fold increase predicted from Eq. (8). This underestimate is reasonable given the assumption that the size of the protein does not change upon denaturation and the assumption that higher virial coefficients can be ignored. In summary, the data are consistent with the idea that the concentration dependence of the calorimetric data arises from nonideality, not from deviation from a two-state process. Tanford predicted this phenomenon in 1961.[34]

Closing Remarks

To observe complete thermal transitions for proteins with high T_m values, experiments are often performed at acid pH. Unfortunately, at acid pH, proteins carry a higher positive charge than at physiological pH, giving rise to nonideal solutions. If the virial coefficients of the native and denatured states are different, the thermodynamic parameters will depend on protein concentration. Note that this concentration dependence occurs even when only the native and denatured states are populated. Therefore, nonideality as well as end-state aggregation and accumulation of intermediates should be considered when apparent deviation from two-state behavior is observed, especially at low pH. Nonideality also should be considered when comparing thermodynamic parameters obtained at low concentration (e.g., CD-detected denaturation), to those obtained at higher concentrations (e.g., nmr-detected amide proton/deuterium exchange).

This work was supported by U.S. Army Medical Research and Material Command under DAMD17-94-J-4270 and by NIH grant GM42501. JCW was partially supported by a GAANN Fellowship from the U.S. Department of Education. We thank C. Lombardo for assistance with ultracentrifugation, D. S. Cohen for assistance with DSC, A. Broadwater and L. Hensley for technical assistance, T. Laue and D. Winzor for helpful discussions, D. Erie and members of the Pielak laboratory for a critique of the manuscript, and the reviewers for their many detailed comments and suggestions.

REFERENCES

1. Cohen, D. S.; Pielak, G. J. Protein Sci 1994, 3, 1253–1260.
2. Privalov, P. L. Adv Protein Chem 1979, 33, 167–241.
3. Sturtevant, J. M. Ann Rev Phys Chem 1987, 38, 463–483.
4. Seemüller, U.; Meier, M.; Ohlsson, K.; Müller, H. P.; Fritz, H. Hoppe Z Physiol Chem 1977, 358, 1105–1107.
5. Bode, W.; Papamokos, E.; Musil, D.; Seemüller, U.; Fritz, H. EMBO J 1986, 5, 813–818.
6. McPhalen, C. A.; James, M. N. Biochemistry 1988, 27, 6582–6598.
7. Hyberts, S. G.; Wagner, G. Biochemistry 1990, 29, 1465–1474.
8. Heinz, D. W.; Priestle, J. P.; Rahuel, J.; Wilson, K. S.; Grütter, M. G. J Mol Biol 1991, 217, 353–371.
9. Hipler, K.; Priestle, J. P.; Rahuel, J.; Grütter, M. G. FEBS Lett 1992, 309, 139–145.
10. Hyberts, S. G.; Goldberg, M. S.; Havel, T. F.; Wagner, G. Protein Sci 1992, 1, 736–751.
11. Bae, S. J.; Sturtevant, J. M. Biophys Chem 1995, 55, 247–252.
12. Jackson, S. E.; Fersht, A. R. Biochemistry 1991, 30, 10428–10435.
13. Jackson, S. E.; Moracci, M.; elMasry, N.; Johnson, C. M.; Fersht, A. Biochemistry 1993, 32, 11259–11269.
14. elMasry, N. F.; Fersht, A. R. Protein Eng 1994, 7, 777–782.
15. Rink, H.; Liersch, M.; Sieber, P.; Meyer, F. Nucleic Acids Res 1984, 12, 6369–6387.
16. Grodberg, J.; Dunn, J. J. J Bacteriol 1988, 170, 1245–1253.
17. Moffatt, B. A.; Studier, F. W. Cell 1987, 49, 221–227.
18. Studier, F. W. J Mol Biol 1991, 219, 37–44.
19. Cohn, E. J.; Edsall, J. T. Proteins, Amino Acids, and Peptides as Ions and Dipolar Ions; Reinhold: New York, 1943.
20. Edelhoch, H. Biochemistry 1967, 6, 1948–1954.
21. Gill, S. C.; von Hippel, P. H. Anal Biochem 1989, 182, 319–326.
22. Liggins, J. R.; Sherman, F.; Mathews, A. J.; Nall, B. T. Biochemistry 1994, 33, 9209–9219.
23. Connelly, P.; Ghosaini, L.; Hu, C. Q.; Kitamura, S.; Tanaka, A.; Sturtevant, J. M. Biochemistry 1991, 30, 1887–1891.
24. Betz, S. F.; Liebman, P. A.; DeGrado, W. F. Biochemistry 1997, 36, 2450–2458.
25. Liu, Y.; Sturtevant, J. M. Biochemistry 1996, 35, 3059–3062.
26. Ptitsyn, O. B.; Birshtein, T. M. Biopolymers 1969, 7, 435–445.
27. Privalov, P. L.; Ptitsyn, O. B. Biopolymers 1969, 8, 559–571.
28. Elwell, M. L.; Schellman, J. A. Biochim Biophys Acta 1977, 494, 367–383.
29. Laue, T. M. Methods Enzymol 1995, 259, 427–452.
30. Johnson, M. L.; Correia, J. J.; Yphantis, D. A.; Halvorson, H. R. Biophys J 1981, 36, 575–588.
31. Marky, L. A.; Breslauer, K. J. Biopolymers 1987, 26, 1601–1620.

Biological Sciences: Biophysics

Patterned library analysis: A method for the quantitative assessment of hypotheses concerning the determinants of protein structure

STEVEN J. LAHR‡, ANNE BROADWATER§, CHARLES W. CARTER, Jr.‡, MARTHA L. COLLIER§, LUCINDA HENSLEY§, JENNIFER C. WALDNER†, GARY J. PIELAK†, and MARSHALL HALL EDGELL§‡*

‡Department of Biochemistry and Biophysics, †Department of Chemistry, and §Department of Microbiology and Immunology, University of North Carolina, Chapel Hill, NC 27599

* Corresponding Author: Dr. Marshall Hall Edgell
Department of Microbiology
CB # 7290, Rm. 741 FLOB
University of North Carolina, NC 27599-7290
Tel: 919-962-0147
Fax: 919-962-8103
marshall@med.unc.edu

Manuscript Information: 27 pages
Word and character count: 165 words in the Abstract
41,458 characters in paper (including title page, abstract, text, acknowledgements, references, figures, figure legends, and tables.

Abbreviation footnote: Single letter code for amino acids: alanine (A), arginine (R), asparagine (N), aspartic acid (D), glutamine (Q), glutamic acid (E), glycine (G), histidine (H), leucine (L), threonine (T), lysine (K), and proline (P)

Data deposition footnote: none

ABSTRACT

Site-directed mutagenesis and combinatorial libraries are powerful tools for providing information about the determinants of protein structure. Here we report two methods that extend their effectiveness. First, we show that resin splitting technology, which allows the construction of arbitrarily complex libraries of degenerate oligonucleotides, can be used to construct more complex protein libraries for hypothesis testing than can be constructed from oligonucleotides limited to degenerate codons. Second, using eglin c as a model protein, we show that regression analysis of activity scores from library data can be used to assess the relative contributions to the specific activity of the amino acids that were varied in the library. The regression parameters derived from the analysis of a 455 member sample from a library wherein four solvent-exposed sites in an α-helix can contain any of nine different amino acids are highly correlated ($P < 0.0001$, $R^2 \geq 0.97$) to the relative helix propensities for those amino acids, as estimated by a variety of biophysical and computational techniques.

Site-directed mutagenesis (1) and combinatorial libraries (2-28) have been used to generate considerable information about the structure-determining elements in proteins. The fraction of variants in combinatorial libraries containing randomized residues (2-11) or residues constrained to be hydrophobic (12-23) that pass some test have been used as semi-quantitative assessments of the role of the targeted residues or motifs. A more hypothesis-oriented approach used degenerate codons to construct binary or hydrophobic-hydrophilic patterns in library variants whose effects on protein structure could then be tested (24-28). However, only a limited number of types of hypotheses can be tested if one uses degenerate codons to introduce variability into a library. We report here two methods to extend the range of hypotheses that can be quantitatively assessed with combinatorial libraries.

The first method involves the use of resin-splitting technology (29,30) to facilitate the construction of arbitrarily complex libraries that are free of the constraints imposed by the genetic code. Libraries can be constructed so that all of their members conform to some hypothesis and members can then be scored by some structurally-related test. As in previous applications, the successful fraction of variants serves as a relative 'score' of the hypothesis, but the arbitrarily complex nature of the hypotheses made possible by split-resin technology extends the range of what can be tested.

The second method involves the use of regression analysis to extend the analytical power of combinatorial library experiments. Appropriate selection of the nature of variants in such a library makes it possible to use regression analysis for quantitative assessment of specific hypotheses and, by averaging over the effects of many factors, to extract accurate information regarding partial effects contributing to protein structure formation. Regression analysis can also be used to assess several competing hypotheses using a single library, in contrast to the approach using the fraction of the library variants that remains active as the metric. Regression analysis provides access to new information by providing a formalism for the quantitative evaluation of the consequences of the effects defined in a hypothesis and a statistical assessment of the degree to which variant behavior can be attributed to them (31-33).

MATERIALS AND METHODS

Reagents. Coomassie® Plus reagent was obtained from Pierce, Suc-Ala-Ala-Pro-Phe-pNA (AAPF) from Bachem, dibromomethane from Aldrich, restriction enzymes and ligase from New England Biolabs, Ni-NTA spin columns from Qiagen, and proteinase K from Qiagen.

Oligonucleotide Synthesis. To synthesize the desired degenerate oligonucleotides, synthesis on a Beckman Oligo 1000 apparatus was interrupted at the positions of interest, the columns opened, the resin removed into an isodense solution (dibromomethane plus 29.4% v/v acetonitrile) to facilitate apportioning. The resin was then allocated into empty Beckman synthesis columns and the appropriate codons added to the growing oligonucleotide chain. The columns were then reopened and the resin from the several columns mixed and returned to another column for continued synthesis. A metal clamp to hold the tops onto the used columns was fabricated to allow their reuse.

Library Construction. A synthetic gene for eglin c was inserted into the pET28a (Novagen) expression vector which adds a his-tag to the N-terminus of eglin to aid purification (34). One PCR primer degenerate at the sites of interest and a second primer, both containing an EarI site on their 3' ends, were used to amplify the entire wild type eglin c template vector (pET28a). The amplified DNA was gel purified, cleaved with EarI, ligated with T4 ligase and transformed into *E. coli* NovaBlue (Novagen). Each library contained any of seven amino acids at four positions (22, 23, 26, and 27; all residue positions are with respect to first codon in wild type eglin c being position 1). The residues at these positions in wild type eglin are R,E,T and L respectively. Six of the seven amino acids were the same in all three libraries (K,Q,E,D,N, and H). One library had P as the seventh amino acid and contained 114 members, another had A and 187 members and the third G and 155 members.

DNA Sequencing. 211 of the variants were sequenced from double stranded DNA prepared by PCR from colonies of variants and were sequenced in the UNC DNA sequencing facility. For these variants sequence was determined for the first 61 of the 76 codons of the his-tagged version of eglin c. The remaining 244 variants were sequenced from single stranded DNA by using the dideoxy-

termination method (Amersham T7 sequanase 2.0 sequencing kit). For these variants sequence was read for codons 10 through 33.

Protein Preparation. To obtain variant proteins for specific activity measurements 1.5 ml cultures were grown in 2XYT medium, induced with 1 mM IPTG for eglin expression at 0.6 OD_{650} and incubated at 37 ºC for 2 hrs. The cultures were then spun at 3000 g in a Beckman GM-3.8 horizontal rotor and the pellets harvested and frozen at -70 ºC. Cell pellets were later thawed at room temperature and resuspended in 1 ml of 50 mM Tris, pH 8.5. An equal volume of lysis buffer (50 mM Tris pH 8.5, 2% Tween 20, 2 mg/ml lysozyme) was added and the mixture allowed to incubate for 20 min at room temperature. Viscosity was reduced by adding 1 M $MgCl_2$ to a final concentration of 10 mM and DNase to a final concentration of 13 units/ml. Debris was removed by centrifugation at 3000 g in a refrigerated centrifuge. 4 M NaCl was added to the supernatant to a final concentration of 300 mM. This solution was added to Qiagen Ni-NTA spin columns prepared as per manufacturers directions, washed 2 times with 600 µl of 50 mM Tris, pH 6.4, 300 mM NaCl and eluted twice with 180 µl of 25 mM citrate, pH 4.5, 300 mM NaCl all by centrifugation at 700 g as per the manufacture's directions. The bulk of the purified protein elutes in the first fraction.

Relative Specific Activity Measurements. These assays were carried out by using a Biomek 2000 robotic liquid handling apparatus. Protein concentrations were determined by mixing 75 µl of sample with 75 µl of Coomassie® Plus reagent in a 96 well plate. After 60 min of color development the optical densities of the wells at 562 or 595 nm were determined by using a Molecular Devices 96 well plate reader. Values from triplicate aliquots were converted to µg/ml of eglin c from a standard curve of purified his-tagged wild type eglin c assayed on the same plate.

Eglin c activity measurements were made by mixing 25 µl of various dilutions of the sample with 40 µl of proteinase K at 0.8 µg/ml in 50 mM Tris, pH 8.5. After a 10 min incubation at room temperature, 40 µl of substrate (AAPF) at 0.6 µg/ml in 175 mM Tris, pH 8.5 was added. After 30 min of color development, the OD_{405} of the sample was determined in a Molecular Devices 96 well plate reader.

Activity (in wild type equivalent μg) was determined by the dilution factor giving rise to 50% of maximal color development referenced to a sample of purified his-tagged wild type eglin at 50 μg/ml assayed on the same plate.

The relative specific activity was calculated by dividing the activity of the variant in wild-type eglin equivalent μg/ml by the protein concentration of the variant in wild-type eglin equivalent μg/ml. The detection limit of sensitivity of this measurement is 0.02 relative specific activity.

Regression Analyses. Regression analyses were carried out with the JMP software package v. 3 (SAS Institute, Inc., Cary, NC).

RESULTS AND DISCUSSION

To explore the utility of using split-resin technology to construct libraries arbitrarily complex patterns we asked whether 'patterned' libraries, in which each library member conforms to some complex pattern or hypothesis, could be used to reproduce a known feature of helix stability, namely, the intrinsic tendency of amino acids to adopt helical dihedral angles (α-helix propensities). We chose eglin c as our model protein because it is a small, 70 residue, protein with both an α-helix and a β-sheet and no cysteines. Its structure is known both from NMR (35) and x-ray crystallography (36), its denaturation thermodynamics is well defined (37,38) and its structural homologue, CI-2, has proven to be a well-mannered protein in folding (39-43) and mutagenesis studies (44-47). We also wanted a protein with a straight-forward activity assay so that we could use high-throughput methods to monitor the consequences of mutation. Eglin c is a proteinase inhibitor that acts by binding so tightly to its target in the Michaelis complex that few eglin molecules make it into the transition state (48). The proteinase binding site is contained within a ten amino acid loop that is on the opposite side of eglin from the α-helix that contains the mutated residues in our libraries (Figure 1). Hence, it seemed plausible that substitutions in the helix might only effect binding via changes in stability.

Percent-Passed Analysis. To determine the feasibility of testing more complex patterns made possible by split-resin technology we constructed three libraries designed to test three hypotheses concerning α-helix propensity. Extensive studies on amino acid propensities in helices (49-55) indicate that variants with hydrophilic amino acids plus alanine substitutions at solvent exposed sites (but not in the N-cap or C-cap) would do well, that variants with hydrophilic amino acids plus proline substitutions at those sites would do poorly, and that variants with hydrophilic amino acids plus glycine would be intermediate. Libraries that can test the effects of adding alanine, proline or glycine to a common set of hydrophilic amino acids cannot be constructed by using degenerate codons. Using split-resin technology we constructed libraries in which each variant in a given library has substitutions at four solvent exposed positions in the eglin α-helix (Figure 1). At each of the four positions a variant could have any of seven

different amino acids. Six of the amino acids (E, K, Q, D, N, and H) are hydrophilic and common to all three libraries. The seventh amino acid is P for one library, G for another and A for the third.

Transformants were selected from each library at random, DNA sequence from the eglin c coding region of each variant was collected to verify consistency with the library design, variant protein was purified by use of the N-terminal his-tag, and the relative specific activity of each variant was measured. We used the percent of each library that was active as the metric for the quality of the library and hence for the hypothesis that it encoded. Since the amount of activity that qualifies a variant as inactive is arbitrary we measured the percent active by using successively higher values of activity as the threshold to determine whether a variant was active or inactive. This statistical device, called a survival curve, is used in many applications. Comparing the survival curves for the three libraries we find that they order as expected (Figure 2); for any activity threshold the proline library has the smallest fraction above threshold, the glycine library a higher fraction, and the alanine library the highest. Thus, for the four eglin helix positions the use of alanine led to the highest "survival rate", supporting the hypothesis that alanine stabilizes α-helices while proline disrupts them.

Regression Analysis: Intrinsic Helix Propensities. Regression analysis tools can be used to extend the capacity to extract information from these libraries. The combined libraries comprise 455 variants whose specific activities span a forty-fold range. The size of this random sample, combined with the variance of their activities affords an opportunity to go beyond the conclusions drawn from a percent-passed analysis and to assess the relative contributions to activity of all of the amino acids that were varied in the libraries. We turned to regression methods because many factors are thought to contribute to α-helix stability. Amino acids have different intrinsic tendencies to form helical dihedral angles (49-55); their sidechains interact differently with the helix macrodipole (56), as well as with each other (57-59); and certain combinations facilitate helix capping (60-65). Regression analysis gives an appropriate evaluation of parameters in a partially correct linear model, as long as the predictor variables are uncorrelated (35). The regression model tested (Eq.1) is that some portion of the response, here eglin

$$\text{Relative Specific Activity} = \Sigma\, k_i * \text{numAA}_i \quad (1)$$

relative specific activity, is a linear sum of effects from amino acids in the varied positions, where the parameter k_i represents the contribution of the i^{th} amino acid type to the activity of a variant and the predictor variable, numAA_i, represents the number of the i^{th} amino acid type in the four substitution sites. That is, numAA_i ranges from 0 to 4 but the sum ($\Sigma\, \text{numAA}_i$) for a given variant must always equal four making this a special kind of regression model, that is, a 'mixture' model in which there is no intercept term (66). Least squares minimization of the difference between predicted and measured relative specific activities generates estimates for the parameters, k_i.

All of the parameters (Table 1) determined from the relative specific activity data for the combined libraries containing 455 variants have P values at least an order of magnitude smaller than that generally taken as a standard for significance ($P < 0.05$). The model as a whole is highly significant ($P < 0.0001$). As expected for a partial model for helix stability, the model does not account for all of the variance seen in the activities of the library members accounting for only 31% of that variance. This observation illustrates one of the attractive features of regression analysis, that is, its capacity to give appropriate values for effects in the model even when other important effects are neglected. For example, in our collection of variants we might expect to find mutations other than those introduced by design but these adventitious mutations do not interfere with our capacity to evaluate the impact of the modeled effects. We estimate that the standard deviation of the mean for the specific activity measurements is ±6% and the irreducible variance in our data, that is, the average difference in activity between variants with the same amino acid composition but different sequence, is 7.9%. This latter value is surprisingly small given the fact that there are other known factors that impact on stability.

Two lines of evidence support the notion that these regression parameters represent helix propensities. First, they accurately reproduce similar estimates for helix propensities obtained by a variety of computational and biophysical methods. Second, they are quite robust, in that they do not vary significantly with the choice of dataset or regression model.

Our regression parameters correlate well (Figure 3A) with the helix propensities values determined from physical measurements in host-guest experiments ($P<0.0001$, $R^2=0.98$ ref. 54 and $P < 0.0001$ $R^2=0.97$ ref. 49). These physical measurements are based on the fraction of model peptides folded and the application of helix-coil transition theory. Our regression parameters also correlate well with propensity values derived from changes in the free energy of denaturation after mutating solvent exposed positions at internal sites in α-helices in globular proteins ($P < 0.0001$ $R^2=0.97$ barnase, ref. 57 and $P < 0.0001$ $R^2=0.98$ T4 lysozyme, ref.52) and with propensities derived from statistical analyses ($P < 0.0001$, $R^2=0.99$, ref. 53).

Although these correlations are as good as that obtained from comparing biophysical helix propensity data from different laboratories [e.g., comparing data from the DeGrado (49) and Baldwin (54) laboratories gives $P < 0.0001$, $R^2 = 0.93$] the correlations are dominated by a single point, the value for proline. However, the correlations between the propensity values for the eight non-proline residues, while lower (Figure 3B), are also as good as the correlation between biophysical data from two different laboratories for those same eight amino acids (Figure 3C). The probability of finding this level of correlation (for the eight amino acids) by chance in a population where there was no correlation between the parameters and activity is less than 1 in 10,000.

How robust are these regression parameters? We initially analyzed a subset of 192 variants using lysate activity instead of specific activity. Although the coefficient of variation of those activity measurements is large (30%), similar regression parameters were obtained (data not shown) as when we used specific activities. Analyzing subsets of the specific activity data shows that there is considerable fluctuation in the regression parameters until the size of the library analyzed reaches 200 to 300 variants (Figure 4). The t-test probability value for the regression parameters for E,K,D and N goes below 5% after analyzing only 30 variants, the parameters for Q and H after 60, for P after 90, for A after 150 and G not until 270 variants are analyzed. If the data are analyzed with a partial model containing only the six hydrophilic residues in common in the three libraries, then the six regression parameters correlate

with an R^2 of 0.96 with those from the model with all nine variant residues. If the subset of 375 variants which have a relative specific activity of greater than 0.3 are analyzed separately, the regression parameters obtained correlate with an R^2 of 0.99 with those obtained from the complete data set and the reduced data set accounts for about the same proportion of the variance.

CONCLUSIONS

Both the percent-passed and regression analysis approaches for utilizing patterned libraries to assess hypotheses concerning the determinants of protein structure reproduce the known effects of amino acid propensity on α-helix stability. However, regression analysis is less stringent in its requirements for use than the percent-passed approach. In the percent-passed approach all of the members of the library must conform to the hypothesis to be tested and, as a consequence, only a single hypothesis can be tested per library. This requirement also places a serious burden on library construction to prevent the inclusion of variants that do not conform to the design criteria. Neither of these constraints applies to regression analysis.

This use of regression analysis to assess the effects of helix propensity might suggest that randomized libraries would be more useful than patterned ones since in our case we might have been able to collect information about all twenty of the amino acids. However, one cannot explore all of the relevant sequence space in mutagenesis experiments. To achieve a useful signal to noise ratio a library must contain a significant number of variants that satisfy the hypothesis if it is to be adequately tested either by percent-passed or by regression analysis (Figure 4). This is particularly true if the analysis is of effects that involve more than one residue such as side-chain interaction effects. In such cases the sample size necessary to contain enough of the hypothesis-testing variants to attain statistically significant results, increases as the degree of patterning decreases.

Our quantitative analysis of patterned libraries extends the traditional anecdotal uses of mutagenesis to test hypotheses. Traditionally, one uses the functionality of variants to confirm the capacity of a hypothesis to predict variant behavior, while unexpected variant behavior suggests the need for modified hypotheses. The percent-passed approach reformulates the former analysis in more quantitative terms. The regression model approach provides access to new information by providing a formalism for quantitative expression of effects in a hypothesis and a statistical assessment of the degree to which variant behavior can be attributed to them. In our case, these proof of principle experiments

produce an independent and reliable index of helix propensities, by virtue of the ability of randomized sampling to average over the effects of potentially confounding effects, correctly revealing partial contributors to helix stability. In addition to quantifying the predictive power of the hypothesis, regression analysis also identifies variants whose behavior deviates from the hypothesis expressed in the model, that is, those with large differences between observed and calculated behavior. These outliers are a rich source of new indications on which the iteration of hypothesis generation, testing and revision depends.

Studies of the effects of amino acids in a model protein are always subject to the idiosyncratic features of that protein. Nevertheless, there are contexts where it is useful to determine the idiosyncratic properties for their own sake. For example, threading algorithms are often designed to utilize idiosyncratic features of the target protein to assess the likelihood that a different sequence is compatible with the target fold. That is, a context is defined for a given residue and probability tables are derived for the amino acids within that idiosyncratic context (67). The results presented here suggest that the analysis of patterned libraries will provide an additional tool for the determination of those probability tables.

Intrinsic α-helix propensity values have been previously derived from several different types of studies. The free energies from various biophysical analyses of peptides (49,50,54) and model proteins (52,57,62) and the pseudo energy terms derived from statistical analyses of the protein database (53) agree remarkably well (55,59). Parameters derived from our analyses of patterned libraries correlate with these previously derived values to the same degree that the previously derived values correlate with each other. This agreement encourages us to believe that our approach can provide quantitative assessments of many hypotheses about the relations between amino acid sequence, stability and structure.

Acknowledgments

We thank Drs. Edward Collins and Robert Bourret for a critical reading of this manuscript. This work was supported by The Department of Defense grant number: DAMD17-94-J-4270 and by National Institutes of Health grants GM-21313 and GM-42501.

References

1. Hutchison, III, C.A.,. Phillips, S., Edgell, M.H., Gillam, S., Janke, P., & Smith, M. (1978) *J. Biol. Chem.* **253** 6551-6560

2. Auld, D.S. & Pielak, G.J.(1991) *Biochemistry* **30** 8684-8690

3. Fredricks, Z.L. & Pielak, G.J.(1993) *Biochem.* 32 929-936

4. Ku, J. & Schultz, P.G.(1995) *Proc. Natl. Acad. Sci. USA* **92** 6552-6556

5. Beasley, J.R. & Pielak, G.J.(1996) *Proteins* **26** 95-107

6. Jeltsch, A., Sobotta, T., & Pingoud, A.(1996) *Protein Eng.* **9** 413-423

7. Petrosino, J.F. & Palzkill, T. (1996) *J. Bacteriol.* **178** 1821-1828

8. Rosenberg, A.H., Griffin, K., Washinton, M.T., Patel, S.S., & Studier, F.W. (1996) *J. Biol. Chem.* **271** 26819-26824

9. Ayabe, T., Takenaka, H., Takenaka, O., Onitsuka, T., Shibata, K., Uesugi, S., & Hamad, M.(1997) *Biochem. Mol. Biol. Int.* **41** 367-375

10. Giese, K.C. & Spudich, J.A. (1997) *Biochemistry* **36** 8465-8473

11. Szmelcman, S., Sassoon, N. & Hofnung, M. (1997) *Protein Sci.* **6** 628-636

12 . Reidhaar-Olsen, J.F., & Sauer, R.T. (1988) *Science* **241** 53-57

13 . Lim, W.A. & Sauer, R.T. (1989) *Nature* **339** 31-36

14 . Hu, J.C., O'Shea, E.K., Kim, P.S., & Sauer, R.T. (1990) *Science* **250** 1400-1403

15 . Lim, W.A. & Sauer, R.T. (1991) *J. Mol. Biol.* **219** 359-376

16 . Sandberg, W.S. & Terwilliger, T.C. (1991) *Proc.Natl.Acad.Sci.USA* **88** 1706-1710

17 . Lim, W.A, Farruggio, D.C, & Sauer, R.T. (1992) *Biochemistry* **31** 4324-4333

18. Baldwin, E.P., Hajiseyedjavadi, O., Baase, W.A., & Matthews, B.W. (1993) *Science* **262** 1715-1718

19. Baldwin, E.P. & Matthews, B.W. (1994) *Curr. Opin. Biotechnol.* **5** 396-402

20. Munson, M, O'Brien, R, Sturtevant, J.M, & Regan, L. (1994) *Protein Sci.* **3** 2015-2022

21. Desjarlis, J.R & Handel, T.M. (1995) *Protein Sci.* **4** 2006-2018

22. Milla, M.E. & Sauer, R.T.(1995) *Biochemistry* **34** 3344-3351

23. Axe, D.D., Foster, N.W., & Fersht, AR. (1996) *Proc.Natl.Acad.Sci.USA* **93** 5590-5594

24. Kamtekar, S, Schiffer, J.M, Xiong, H., Babik, J.M, & Hecht, M.H. (1993) *Science* **262** 1680-1685

25. West, M.W & Hecht, M.H. (1995) *Protein Sci.* **4** 2032-2039

26. Riddle, D.S., Santiago, J.V., Bray-Hall, S.T., Doshi, N., Grantcharova, V.P., Yi, Q., & Baker, D. (1997) *Nat. Struct. Biol.* **4** 805-809

27. Rojas, N.R., Kamtekar, S., Simons, C.T., McLean, J.E., Vogel, K.M., Sipro, T.G., Farid, R.S., & Hecht, M.H. (1997) *Protein Sci.* **12** 2512-2524

28. Roy, S., Helmer, K.J., & Hecht, M.H. (1997) *Fold Des.* **2** 89-92

29. Glaser, SM., Yelton, DE., & Huse, WD. (1992) *J. Immun.* **149** 3903-3913

30. Huse, W.D., Yelton, D.E., & Glaser, S.M. (1993) *Int. Rev. Immunol.* **10** 129-137

31. Doublié, S.S., Gilmore, X., Bricogne, C.J., & Carter, C.W., Jr. (1994) *Acta Crystallographica* **A50** 164-182

32. Carter, C.W.,Jr., Doublié, S.S., & Coleman, D.E. (1994) *J. Mol. Biol.* **238** 346-365

33. Neter, J. Kutner, M.H., Nachtsheim, C.J., & Wasserman, W. 1996 *Applied Linear Statistical Models* 4th ed. p. 285 New York:McGraw-Hill

34. Waldner, J.C., Lahr, S., Edgell, M.H. & Pielak, G.J. (1998) *Anal. Biochem.* **263** 116-118

35. Hyberts, S.G., Goldberg, M.S., Havel, T.F. & Wagner, G. (1992) *Protein Sci.* **1** 736-751

36. McPahlen, C.A.& James, M.N.G. (1988) *Biochemsitry* **27** 6852-6598

37. Bae, S.J. & Sturdevant, J.M. (1995) *Biophys. Chem.* **55** 247-252

38. Waldner, J. C., Lahr, S., Edgell,M.H. & Pielak,G. J. (1999) *Biopolymers* **49** 471-479

39. Jackson, S.E. & Fersht, A.R. (1991) *Biochemsitry* **30** 10428-10435

40. Jackson, S.E. & Fersht, A.R. (1991) *Biochemsitry* **30** 10436-10443

41. Fersht, A.R, Itzhaki, L.S., el Masry, N.F., Matthews, J.M. & Otzen, D.E. (1994) *Proc.Nat.Acad.Sci. USA* **91** 10426-10229

42. Itzhaki, L.S., Otzen, D.E. & Fersht, A.R. (1995) *J. Mol. Biol.* **254** 260-288

43. Itzhaki, L.s., Neira, J.L., Ruiz-Sanz, J., de Prat Gay, G. & Fersht, A.R. (1995) *J. Mol. Biol.* **254** 289-304

44. Harpaz, Y., Elmasry, N., Fersht, A.R.& Henrick, K. (1994) *Proc Nat. Acad. Sci. USA* **91** 311-315

45. Jackson, S.E. & Fersht, A.R. (1994) *Biochemsitry* **33** 13880-13887

46. Chen, Y.Y. & Fersht, A.R. (1994) *FEBS Lett* **347** 304-309

47. elMasry, N.F. & Fersht, A.R. (1994) *Protein Eng.* 7 777-782

48. Shaw, G.L., Davis, B., Keeler, J., & Fersht, A.R. (1995) *Biochemistry* **34**:2225-2233

49. O'Neil, K & DeGrado, W.F. (1990) *Science* **250** 246-250

50. Lyu, P.C, Liff, M.I, Marky, L.A, & Kallenbach, N.R. (1990) *Science* **250** 669-673

51. Hermans, J., Anderson, A. & Yun, R.H. (1992) *Biochem.* **31** 5646-5653

52. Blaber, M., Zhang, X., & Matthews, B.W. (1993) *Science* **260** 1637-1640

53. Munoz, V & Serrano, L. (1994) *Proteins* **20** 301-311

54. Chakrabartty, A., Kortemme, T. & Baldwin, R.L (1994). *Protein Sci.* **3** 843-852

55. Myers, J.K., Pace, C.N. & Scholz, J.M. (1997) *Proc. Nat.Acad. Sci. USA* **94** 2833-2837

56. Shoemaker, K.R., Kim, P.S., York, E.J., Stewart, J.M. & Baldwin, R.L. (1987) *Nature* **326** 563-567

57. Horovitz, A, Matthews, J.M, & Fersht, A.R. (1992) *J.Mol.Biol.* **227** 560-568

58. Munoz, V. & Serrano, L. (1994) *Nature Struc. Biol.* **1** 399-409

59. Munoz, V. & Serrano, L. (1995) *J. Mol. Biol.* **245** 275-296

60. Presta, E.T. & Rose, G.D. (1988) *Science* **240** 1632-1641

61. Richardson, J.S. & Richardson, D.C. (1988) *Science* **240** 1648-1652

62. Serrano, L., Sancho, J., Hirshberg, M., & Fersht, A.R. (1992) *J.Mol.Biol.* **227** 544-559

63. Harper, E.T. & Rose, G.D. (1993) *Biochem.* **32** 7605-7609

64. Doig, A.J. & Baldwin, R.L. (1995) *Protein Sci.* 4 1325-1336

65. Aurora, R. & Rose, G.D. (1998) *Protein Sci.* 7 21-35

66. Marquardt, D.W. & Snee, R.D. (1974) *Technometrics* **16** 533-537

67. Bowie, JU., Luthy, R. & Eisenberg, D. *Science* **253** 164-170 (1991)

Figure Legends

Figure 1. Ribbon diagram of eglin c. The proteinase binding site in eglin c is contained within the ten amino acid loop in the upper left of the diagram. The solvent exposed residues in the α-helix that are varied in the libraries (R22, E23, T26 and L27) are shown in stick-ball format.

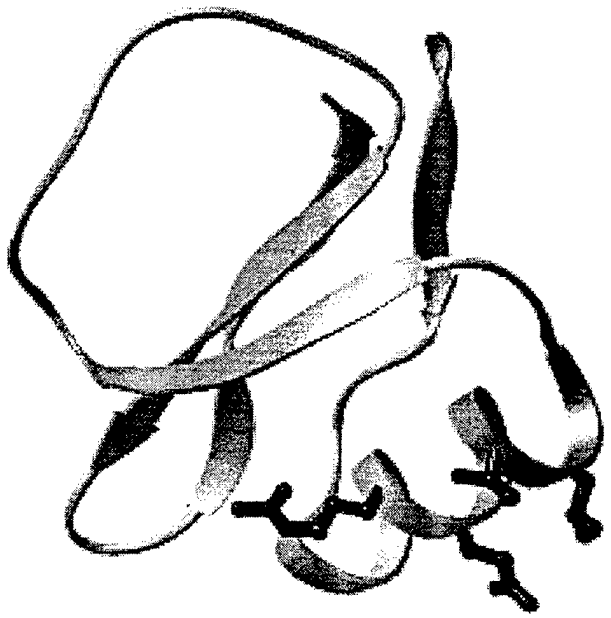
Figure 2. Percentage of variants in the libraries that have relative specific activities above a given threshold. The wild type sequence is defined to have a relative specific activity of 1.00. The library with the six common hydrophilic residues and alanine, 187 members, is indicated with a (❑); the hydrophilic residues plus glycine library, 154 members, (○); and hydrophilic residues plus proline, 114 members,(Δ).

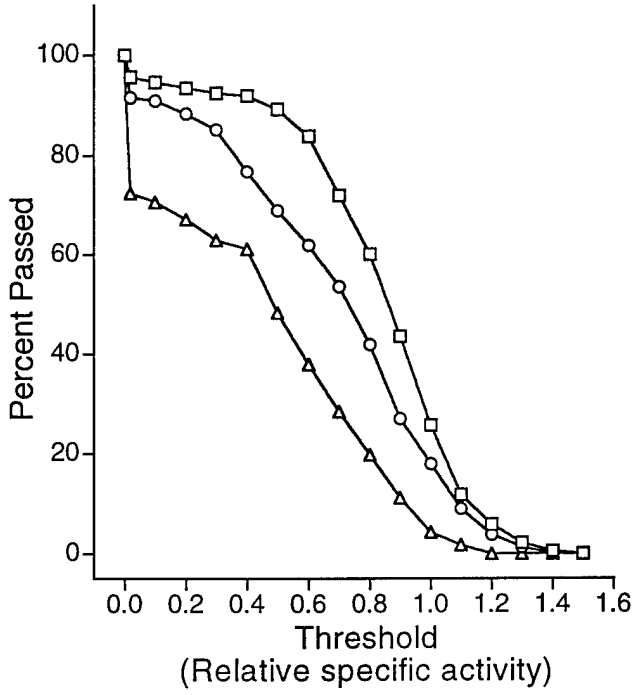
Figure 3. Correlation between intrinsic helix propensity values for amino acids determined by biophysical methods and by regression analysis of a patterned library. Panel A compares the regression analysis parameters of all nine amino acids that were variant in our library with α-helical propensity values derived from biophysical measurements (guest-host experiments in alanine peptides) on peptides in aqueous solution (54). Panel B compares regression parameter values of the variant amino acids minus proline with α-helical propensity values derived from biophysical measurements (54). Panel C compares α-helical propensity values derived from biophysical measurements in two different laboratories [Baldwin (54) and DeGrado (49)]. The solid lines have been fit by least squares and the dotted curves represent the 95% confidence limits of the fit.

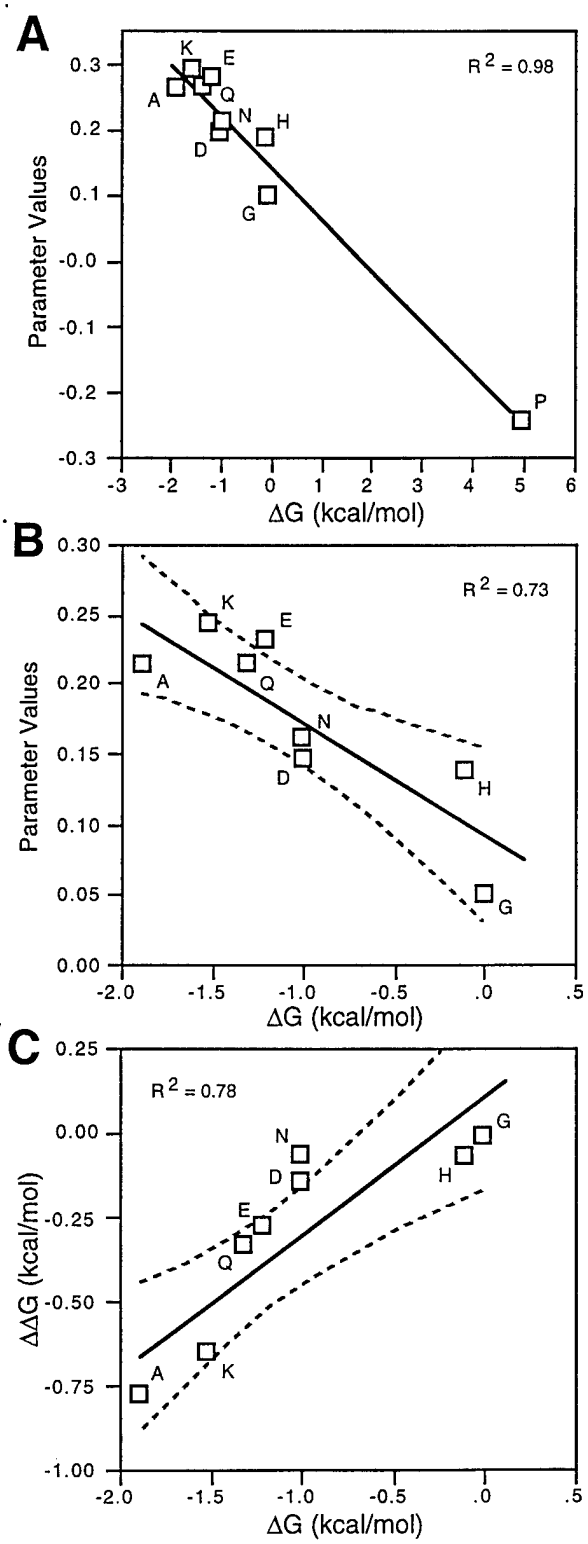
Figure 4. Regression parameters versus library size. Regression parameters for the effects of nine variant amino acids were calculated for subsets of the libraries to determine how they varied as a function of library size. Five of the parameters covering the range of behaviors are shown. The subsets contained equal numbers of variants from the three libraries until a subset size of 360 at which point the

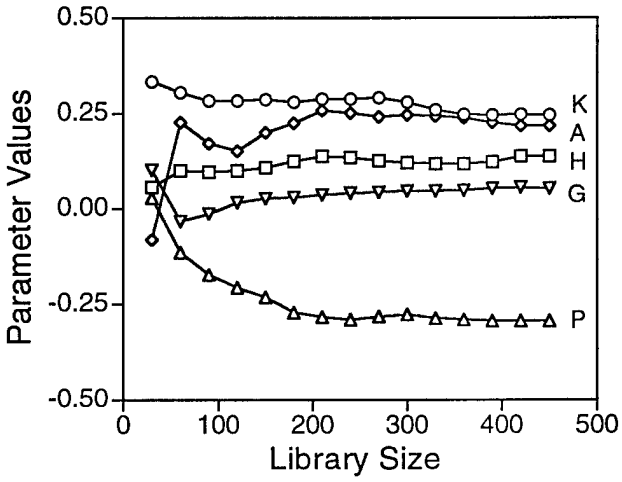
120 variants from the proline library were exhausted. The variants in the subsets were chosen in the order they were picked from the transformation plates.

Illustrations (Full Size)




100
80
60
40
20
0
Percent Passed
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Threshold
(Relative specific activity)


A
R 2 = 0.98
Parameter Values
ΔG (kcal/mol)
K
E
Q
A
N
H
D
G
P
B
R 2 = 0.73
Parameter Values
ΔG (kcal/mol)
C
R 2 = 0.78
ΔΔG (kcal/mol)
ΔG (kcal/mol)


0.50
0.25
0.00
-0.25
-0.50
Parameter Values
0
100
200
300
400
500
Library Size
K
A
H
G
P

Tables.

Amino Acid	k_i	StdErr	P>|t|
K	0.248	0.020	< 0.0001
E	0.233	0.016	< 0.0001
A	0.217	0.034	< 0.0001
Q	0.217	0.016	< 0.0001
N	0.165	0.019	< 0.0001
D	0.150	0.019	< 0.0001
H	0.139	0.016	< 0.0001
G	0.052	0.018	0.0051
P	-0.295	0.043	< 0.0001

Table 1. Regression Analysis Parameters for Effects of Amino Acid Composition in Four Helix Sites on Activity. The parameter (k_i) represent the fraction of the relative specific activity contributed by that amino acid and hence are unitless. StdErr estimates the standard deviation of the distribution of the parameter estimate. P>|t| is the probability of getting an even greater t statistic given the hypothesis that the parameter is zero, that is, that there is no correlation between the particular amino acid composition and the relative specific activity.